6-2-2023

Signal Processing Based Pathological Voice Detection Techniques

Rumana Islam
University of Windsor

Signal Processing Based Pathological Voice Detection Techniques

By

Rumana Islam

A Dissertation
Submitted to the Faculty of Graduate Studies
through the Department of Electrical and Computer Engineering
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy
at the University of Windsor

Windsor, Ontario, Canada

Signal Processing Based Pathological Voice Detection Techniques

By

Rumana Islam

APPROVED BY:

_____

A. Elmaghraby, External Examiner
University of Louisville

_____

L. Rueda
School of Computer Science

_____

M. Khalid
Department of Electrical and Computer Engineering

_____

K. E. Tepe
Department of Electrical and Computer Engineering

_____

E. Abdel-Raheem, Advisor
Department of Electrical and Computer Engineering

March 23, 2023

DECLARATION OF CO-AUTHORSHIP AND PREVIOUS PUBLICATION

I. Co-Authorship

I hereby declare that this thesis incorporates the contributions of joint research. Chapters 2-6 of this thesis were completed under the supervision of Dr. Esam Abdel-Raheem. In all cases, the author performed key ideas, survey works, experimental simulations, methodology, data analysis, graphical investigations, and writings. The contribution of my supervisor, Dr. Esam Abdel-Raheem was providing comments and valuable feedback on the refinement of ideas, editing the manuscript, and advising on the selection of peer-reviewed journals for possible publications. This dissertation also includes the outcome of publications co-authored by Dr. Mohammed Tarique. His contribution was providing feedback, refining methodology, and editing the manuscript.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis, and have obtained written permission from each of the co-author(s) to include the above material(s) in my thesis.

I certify that, with the above qualification, this thesis, and the research to which it refers, is the product of my own work.

II. Previous Publications

This dissertation partly includes five original papers that have been published, or to be submitted for publication in peer-reviewed journals as follows:

Dissertation Chapter	Publication Title	Publication status
Part of Chapter 2	R. Islam, M. Tarique, and E. Abdel-Raheem, "A Survey on Signal Processing Based Pathological Voice Detection Techniques," *IEEE Access*, vol. 8, pp. 66749 – 66776, April 2020.	Published

Part of Chapter 3	Rumana Islam, E. Abdel-Raheem, and Mohammed Tarique, "Voice Pathology Detection using Convolutional Neural Network with Electroglottographic (EGG) and Speech Signals," *Computer Methods and Programs in Biomedicine Update*, *Elsevier*, vol. 2, 100074, pp. 1–13, October 2022.	Published
Part of Chapter 4	Rumana Islam, E. Abdel-Raheem, and Mohammed Tarique, "A Novel Pathological Voice Identification Technique through Simulated Cochlear Implant Processing Systems," *Applied Science,* vol. 12, no. 5, pp. 1-21, February 2022.	Published
Part of Chapter 5	Rumana Islam, E. Abdel-Raheem, and Mohammed Tarique, "A study of using cough sounds and deep neural networks for the early detection of COVID-19," *Biomedical Engineering Advances, Elsevier*, vol. 3, 100025, pp. 1–12, June 2022.	Published
Part of Chapter 6	Rumana Islam, E. Abdel-Raheem, and Mohammed Tarique, "Cochleagram to Recognize Dysphonic Voice: An Auditory Perceptual Analysis."	To be submitted

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as a graduate student at the University of Windsor.

III. General

I declare that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices.

Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

Voice pathology is not only circumscribed by voice impairment or speech disorder. Pathological voice is also a biomarker of neuropsychiatric and neurocognitive diseases, including physical and muscular conditions. Alzheimer's, Parkinson's, Schizophrenia, ASD (Autism spectrum disorder), oral/lung cancer, depression, and asthma strongly correlate with voice disability.

The physicians' current endoscopic procedures to detect pathological voices are painful for the patients. Clinical invasive diagnostic procedures, for example, laryngoscopy, laryngeal electromyography, stroboscopy, etc., require high-level expertise; they are expensive and time-consuming. This research focuses on establishing automated voice signal processing-based noninvasive procedures to identify pathological voices with objective diagnostics on top of subjective assessment. Several quick computerized digital signal processing-based techniques are implemented that require no extensive training/expensive equipment. Being noninvasive, they do not traumatize the patients. Also, can evaluate structural, neurological, and respiratory voice disorder. An extensive temporal, spectral, acoustical, wavelet domain based feature analysis is also performed to enhance the current understanding of pathological voice.

DEDICATION

I am dedicating this thesis to the memory of my beloved parents with a unique feeling of love, gratitude, and respect. Their endless love and lessons inspire me to be dedicated to achieve my goal.

ACKNOWLEDGEMENTS

All praises belong to Almighty Allah (SWT) for His generosity in guiding us with unlimited support. I want to express my deep and sincere gratitude to my honorable research supervisor Dr. Esam Abdel-Raheem for his continuous support, encouragement, and guidance throughout my Ph.D. study. His academic support, patience, and motivation inspired me constantly during my journey. I sincerely thank all respectful Ph.D. committee members, Dr. Mohamed Khalid, Dr. Kemal Tepe, and Dr. Luis Rueda, for their valuable comments and thoughtful feedback during my Ph.D. seminars. I appreciate my husband encouraging me to finish my Ph.D. at the University of Windsor. I am grateful for his caring, inspiration, and academic advice. Finally, I would like to express my love and thanks to those behind this journey, my responsible son Ayman Sabih and lovely daughter Aribah Tasnim for their exceptional understanding and patience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS/SYMBOLS

AI	Artificial Intelligence
ANN	Artificial neural network
APR	Ratio of aperiodic/periodic components in cepstral energy
APQ	Amplitude perturbation quotient
ARDS	Acute respiratory distress syndrome
AR-HMM	Autoregressive higher-order HMM
ASR	Automatic speech recognition
AVPD	Arabian voice pathology database
BH	Bhattacharyya distance
CDBN	Convolutional deep belief network
CEP	Cepstral coefficients
CHNR	Cepstral based HNR
CNN	Convolutional neural network
COMPARE	Computational Paralinguistic Challenge
COPD	Chronic obstructive pulmonary diseases
CPP	Cepstral peak prominence
CSL	Computerized speech laboratory
CT	Computerized tomography
CWT	Continuous wavelet transform
DBN	Deep belief network
DCE	Delta cepstral
DFT	Discrete fourier transform
DH	Degree of hoarseness
DNN	Deep Neural Network
DNRNN	Dense Net Recurrent Neural Network
DPP	Dissimilarities in the surfaces of the pitch pulses
DUV	Degree of unvoiceness
DWT	Discrete wavelet transform
EE	Emotional expression

EER	Equal error rate
eGeMAPS	Geneva Minimalistic Acoustic Parameter Set
EGG	Electroglottographic
F0SD	Fundamental frequency and its variations
FFNN	Feed forward neural network
FFT	Fast fourier transform
FIR	Finite Impulse Response
GERD	Gastroesophageal reflux disease
GMM	Gaussian mixture model
GMM-SVM	Gaussian Mixture Model Support Vector Machine
GMM-UBM	GMM-Universal background model
GNE	Glottal-to-noise excitation ratio
GNN	Graph neural network
GNR	Glottal to noise ratio
GQ	Glottal quotients
GRBS	Grade Roughness Breathiness Strain
GRNN	Generalized regression neural network
H1H2	First and second harmonic
HLAC	Higher-order local autocorrelation
HNR	Harmonic to noise ratio
HTK	Hidden Markov model tool kit (HTK)
IIR	All-pole infinite impulse response
JIT	Jitter
KL	Kullback Leibler
KL-MCS	Classical KL
LDA	Linear discriminant analysis
LEMG	Laryngeal electromyography
LFCC	Linear frequency cepstral coefficient
LP	Linear predictor
LSTM	Long short-term memory
LTI	Linear time-invariant system

LPC	Linear Predictive Coding
MAP	Maximum a posterior probability algorithm
MDVP	Multi-dimensional Voice Program
MEEI	Massachusetts Eye and Ear Infirmary
MFCC	Mel-frequency Cepstral Coefficient
MLPNN	Multilayer perceptron neural network
MPT	Maximum phonation time
MRI	Magnetic resonance imaging
NOVA	Analysis of variance
NHR	Noise-to-Harmonic ratio
NNE	Normalized noise energy
NNR	Noise energy to total energy ratio
PA	Pitch amplitude
PCA	Principal component analysis
PDF	Probability density function
PECM	Ratio of cepstral energies
PLP	Perceptual linear prediction
PNN	Probabilistic neural network
PPQ	Pitch perturbation quotients
PSD	Power Spectral Density
RAP	Relative Average Perturbation
RASTA-PLP	Relative spectral transform – PLP
RBF	Radial basis function
RBM	Restricted Boltzmann machines
RBFNN	Radial Basis Functional Neural Networks
RF	Random Forest
RNN	Recurrent Neural Network
ROC	Receiver operating Characteristic
RT-PCR	Reverse transcription-polymerase chain reaction
SIR	Spectral Flatness of the Residue Signal
SOP	Standard operating procedure (SOP)

SPI	Soft phonation index
SVD	Saarbrucken Voice Database
SVM	Support vector machine
STE	Short-time energy
STFT	Short-time Fourier transform
TEO	Teager energy operator
VTI	Voice turbulence index
WCEP	Weighted cepstral
WDCEP	Weighted delta cepstral
WHO	World Health Organization
ZCR	Zero crossing rate

CHAPTER 1
INTRODUCTION

Voice communication is an integral part of our personal and professional life. However, that can be affected by several barriers. Speech impairment due to voice disability is one of them. Approximately 9.4 million adults have problems using voices that last for one week or more. According to a recent National Center for Education Statistics report, about 20% of children and youth in 3-21 years suffer from voice disability [1]. Voice disability occurs mainly from a disruption in the human voice generation system.

1.1 Human Voice Generation System

The human voice generation system mainly consists of the lungs, larynx, and vocal tract [2] , as shown in Fig. 1.1. During voice generation, we inhale air by expanding the rib cage surrounding the lungs. Then, we expel air from the lungs by lowering the diaphragm located at the bottom of the lungs. We maintain a steady flow of air by controlling the muscles around the rib cage depending on the length of the sentence or phrase. This action causes air to rush through the vocal trachea to the epiglottis. The larynx is the most complicated part of the human voice generation system. It consists of cartilages, muscles, and ligaments. The primary purpose of the larynx is to control vocal folds, which include two masses stretched between the front and back of the larynx. A slit-like orifice called glottis exists between the two masses.

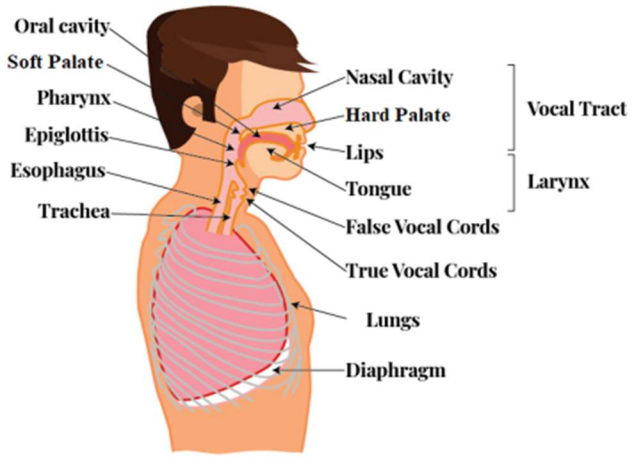


Figure 1.1. The main components of the human voice generation system [2].

During normal conditions, vocal folds are in a state called 'breathing'. Under the breathing state, the vocal fold masses are relaxed, and the glottis is opened. The air from the lungs flows through the glottis without much obstruction, and no vocal fold vibration occurs. During voice generation, vocal folds can be in two states, namely 'unvoiced' and 'voiced'. Under the unvoiced condition, the vocal folds come closer and generate turbulence by themselves. While under voiced condition (i.e., during the generation of a vowel), the vocal folds come significantly closer, become more tensed, and partially close the glottis. The partially closed glottis and increased tension cause oscillation of the vocal folds. The air stream from the lungs is interrupted by the vocal cords, and a quasi-periodic pressure wave is generated. The impulses of this pressure wave are called pitch, and the pressure's frequency is called pitch frequency. The masses of the larynx adjust the length and tension of vocal folds to 'fine-tune' pitch and tone. The articulators (i.e., tongue, palate, cheek, and lips) articulate and filter the sound emanating from the larynx. The vocal fold and articulators produce highly intricate sounds.

1.2 Problem Statement

Voice is a kind of biomedical signal that refers to all possible sounds generated by the human voice generation system, such as vowels, continuous speech, counting numbers, breathing, and coughing sounds. The production of voice necessitates the cooperation of multiple organs, namely (a) the nervous system, (b) the respiratory system, and (c) the vocal cords [3]. The nervous system coordinates the operation of various organs and tissues to generate voice. The respiratory system provides the energy to the vocal process through the lungs and trachea. The vocal cords and vocal tracts act as vibrators and resonators, respectively. A pathological voice is generated when the diseases affect the system/systems mentioned above directly or indirectly.

The causes of all kinds of voice disorders are still unknown. However, calluses on the vocal cords, swelling or bumps like blisters on the vocal cords, vocal cord paralysis, vocal cord shutting, and spasmodic dysphonia are the leading causes of voice disability. Other reasons include hearing loss, neurological disorders, brain injury, intellectual disability, drug abuse, and malfunction of the human voice generation system. In addition, people may encounter temporary voice disorders due to allergies, large tonsils, smoking-related illnesses, respiratory infections, and poor voice habits.

Voice pathology detection is the first crucial step for correctly characterizing and managing voice disorders. Both invasive and non-invasive methods are used for detecting voice pathologies. In invasive techniques, physicians insert probes into the mouth using an endoscopic procedure. Laryngoscopy [4], stroboscopy [5], and laryngeal electromyography [6] are examples of such practices. These approaches are expensive, require specialization, and are invasive. The morbidity rate of the diseases affecting the vocal cords and vocal tracts is so high that complementary automated, objective, and noninvasive diagnostics are imperative.

In the context of the above facts, in non-invasive methods, voice pathology is detected using voice signal processing [7]-[8] techniques. The development of computerized voice analysis algorithms is promising to provide state-of-the-art results for disease detection and monitoring in current clinical settings. These methods involve three significant steps, namely

(a) Voice samples collection and analysis,

(b) Features extraction, and

(c) Classification.

Voice samples are collected in a sound environment. Then, the samples are analyzed, and voice features are extracted. The final step is to classify voice samples into control (i.e., healthy) and pathological. A classifier is commonly used for this purpose.

A literature survey shows that several classifier algorithms have been popularly used for voice pathology detection. The published results demonstrate that the classification accuracy mainly depends on the classifier algorithms and voice features [9]-[10] used by the classifiers. Recently, deep learning algorithms have drawn considerable attention from researchers in this field. It has been shown in [11]-[14] that deep learning algorithms can play an essential role in voice pathology detection as they provide higher accuracies.

1.3 Research Objectives

The biomedical value of voice is less maneuvered than biometric applications, such as speech recognition or speaker identification. The variation in sounds that can be produced both among and between individuals is virtually infinite, conveying psychological and physiological information about the speaker. Significant progress has been made in

understanding the process of normal vocalization and how various speech pathologies can arise. One growing area is an audiometric analysis of voiced sounds. More recently, advanced algorithms such as machine learning and deep learning have further advanced our understanding of the human voice.

This research intends to use signal processing techniques and artificial intelligence algorithms to classify and diagnose human voices among individuals with and without pathology. In addition, these methods may prove helpful in many applications, such as in detecting and classifying voice disorders (structural and neurological) through voice quality assessment and in disease severity prediction for pathological voices.

1.4 Methodologies and Research Approaches

Voice sample analysis based on noninvasive pathological diagnosis contains the following significant steps.

1.4.1 Voice Acquisition/Recording

This is the first step of computerized voice analysis for disease detection and monitoring. The voice acquisition technique is of great importance for pathological voice analysis since the quality of voice significantly impacts the performance of voice analysis. The influence of SNR (signal-to-noise ratio) and sampling rates are two key factors that require attention. A pathological voice often shows signs in high-frequency bands [12], coinciding with the frequency range of numerous noises. Hence, the influence of environmental noise should be researched further. The required sample collection criterion [15] suggests that the sampling frequency should not be less than 16 kHz and that the voice recording should be conducted in a sound-treated room. Also, the mouth-to-microphone distance should be constant and less than 10 cm (preferably 3-4 cm). There are several publicly available databases, for example, Saarbrucken Voice Database (SVD), Massachusetts Eye & Ear Infirmary (MEEI), Arabian voice pathology database (AVPD), etc., that the researchers are effectively using to design noninvasive algorithms for pathological voice diagnostics [11].

1.4.2 Selection of Voice Samples

Sustained vowels are commonly used for pathological voice analysis for 2-3 seconds to identify the phonation activity of the voice. During vowel sound generation, a speaker can maintain a steady frequency and amplitude at a comfortable level [15]. Moreover, it is free

of articulatory and other linguistic confounds that often exist with everyday speech tasks such as sentences and running speech. Voices of laryngeal diseases, such as vocal fold polyps, adductor spasmodic dysphonia, keratosis leukoplakia, vocal nodules, and vocal fold paralysis, can be investigated from the vowel samples. Some other voice samples that can be considered are counting numbers (say, 1-10), sentences, etc. Physiological voice disorders that result from alterations of laryngeal, respiratory, or vocal tract mechanisms are better revealed with these samples. Breathing and coughing sounds are good indicators for investigating vocal issues related to respiratory disorders like pneumonia, chronic obstructive pulmonary diseases (COPD), bronchitis, COVID-19, etc. [11]. The voice generation systems for males and females are structurally different. Male and female larynges vary in size, vocal fold membranous length, the elasticity of vocal fold tissues, and pre-phonatory glottal shapes [16]-[18]. Hence, the male and female voice samples must be investigated separately to provide an unbiased classification accuracy, as demonstrated in this study.

1.4.3 Design of Signal Processing Algorithms for Voice Samples

Various kinds of features can be exploited to differentiate/classify voice samples for computerized voice analysis. Not all features perform equally well with different classifier algorithms [10]. It is a unique design criterion exercised by biomedical researchers. Based on the broader definition of domains, voice features can be categorized into the time-domain, frequency-domain, and wavelet-domain.

Time-domain features are those features that can be measured based on temporal voice samples. That includes short-time energy (STE), zero-crossing rate (ZCR), short-time entropy of energy, relative average perturbation (RAP), etc. [19].

Frequency-domain features are those features that can be extracted from voice spectrums/short-time Fourier transform (STFT). The formants, Mel-frequency cepstral coefficients (MFCCs), spectral centroid, spectral entropy, spectral flux, spectral roll-offs, feature harmonics, etc. [19] are frequency-domain features.

Wavelet-domain features include continuous wavelet transform (CWT) and discrete wavelet transform (DWT) coefficients. These wavelet-domain features are also good indicators of identifying voice disorders [8]. Noise measurements to describe the

breathiness, hoarseness, and hypernasality using the features harmonics-to-noise ratio (HNR), normalized noise energy (NNE), voice turbulence index (VTI), soft phonation index (SFI), glottal-to-noise excitation ratio (GNE), and glottal quotients (i.e., GQ open and GQ closed), etc. can be investigated [20]-[22] to diagnose voice pathology. Considering the vocal tract as a linear time-invariant system (LTI), different formants, F1, F2, F3, etc., can be estimated to characterize vocal tracts for identifying pathological voices. The tongue, nose cavity, and oral cavity determine the shape of the vocal tract. Pathologies affecting these three elements result in unnatural forms of the vocal tracts that are well reflected through formants analysis [8].

1.4.4 Design of Classifier Algorithms

After extracting suitable voice features from voice samples, the final step is to design a robust classifier to identify and classify pathological voice samples from healthy ones. A discriminative single feature or feature vector can be employed depending on the design criterion. Different machine learning algorithms can be implemented for this purpose, for example, support vector machine (SVM), k-nearest neighbor (kNN), logistic regression, etc. Also, artificial neural network (ANN) based algorithms can be used. The deep neural network (DNN) based algorithms, for example, convolutional neural networks (CNNs), can be designed to solve the classification problems with better accuracy. Large data samples are always suggested to achieve the best accuracy for the DNN. The transfer learning approach can be employed to overcome the small data sample size limitation. Following this approach, a pre-trained CNN is used to identify the lower-level features; consequently, the network is trained to learn the upper-level features from the given data. VGG16, AlexNet, GoogleNet, ResNet, etc., fall in this category showing better performance with limited data samples. Also, a gammatone filter can be designed for voice pathology identification through the signal processing steps involved in the cochlear implant device. This method removes the necessity of feature extraction and selection as used in the conventional approach [22].

1.5 Research Contributions

The major contributions of this dissertation can be summarized as:

- Two CNN based algorithms are proposed to process the raw EGG and voice signals to detect and categorize the pathological voices into pathology types. Considering sustained vowel (‘/a/’) samples as the raw inputs, the proposed system can identify the contribution of EGG and voice signals to identify and classify the three major voice pathologies, namely dysphonia, laryngitis, and vocal fold polyps affecting the general population. The proposed system can extract discriminative features from raw audio samples as opposed to conventional algorithms. Hence it is much faster than most other feature-based systems requiring a special type of General Processing Unit (GPU) to overcome the computation burden
- A novel pathological voice identification system considering the biological process of speech perception is developed to identify the laryngeal voice disorder. The merit of the proposed system is that it eliminates the need for feature extraction from speech samples. An optimized gammatone filter bank is also designed to divide the speech signal into several channels for facilitating the classification process as opposed to a conventional bandpass filter bank. The center frequencies and bandwidths of the gammatone filter bank are designed to mimic the cochlear vibration pattern of the human ear. The performance measure shows significant improvement in terms of the F1-score.
- A low-cost, quick, and easily accessible COVID-19 detection system is proposed by employing deep learning with coughing sound samples of healthy and COVID subjects. The time, frequency, and mixed domain feature vectors extracted from the coughing sound samples are investigated to demonstrate their efficacy in identifying COVID voice. Statistical diagnostic performance measures are significantly high with the frequency domain feature vectors.
- A pre- trained CNN, VGG 16, is employed to devise a system model that can detect dysphonic voice from both sustained vowel (‘/a/’) and speech samples efficiently. A discriminative spectral image, cochleagram is generated from dysphonic and healthy voice samples to unveil their detailed spectral contents. Also, female and male voice samples are considered separately to eliminate gender bias in the detection algorithm. A significant improvement in performance is achieved

considering the transfer learning approach with VGG-16. The limitation of sparse pathological data samples is overcome with the designed model.

1.6 Organization of Dissertation

The motivation of Chapter 2 is to address the need for non-invasive signal processing techniques to detect voice disability in the general population. The first few Sections present background information, including causes of voice disability, current procedures, and practices, experimental setup to include some voice features analysis, and classifiers. Also, comprehensive literature survey work on voice disability detection algorithms is provided. The issues and challenges related to the selection of voice features and classifier algorithms have been addressed at the end of Chapter 2.

Chapter 3 introduces the contribution of EGG and voice signals to identify and classify the three most common voice pathologies.

Chapter 4 presents the design of a noninvasive pathological voice identification system employing a signal processing-based technique through a simulated cochlear implant processing system.

Chapter 5 proposes a low-cost, quick, and easily accessible COVID-19 detection system by employing deep learning with coughing sound samples of healthy and COVID subjects.

In Chapter 6, a pre- trained CNN, VGG 16, is employed to devise a system model that can detect dysphonic voice from both sustained vowel (‘/a/’) and speech samples efficiently.

Finally, the dissertation is concluded with future directions in Chapter 7.

CHAPTER 2
EXPERIMENTAL SETUP AND LITERATURE SURVEY

Researchers and practitioners have suggested voice sample-based noninvasive pathology diagnosis techniques. The significant steps followed by these techniques are (a) extracting voice features from voice samples and (b) discriminating pathological voices from normal voices using a classifier algorithm. However, there is no consensus on the voice feature and classifier algorithm that can provide the best accuracy in screening for voice disability. Moreover, some techniques use multiple voice features and multiple classifiers to ensure high reliability. This Chapter presents a comprehensive survey of signal processing-based pathological voice detection techniques to address the above issues. Sections: 2.1-2.4 of this Chapter offer background information, including causes of voice disability, current procedures, practices, the experimental setup for voice features analysis, and classifiers. The later Section: 2.5, presents a comprehensive survey work on voice disability detection techniques [23]. Section: 2.6 addresses the issues and challenges related to these techniques. Finally, this Chapter is concluded with Section: 2.7.

2.1 Medical Conditions of Voice Pathology

Speech pathologists have related certain medical conditions to voice disability. These medical conditions include asthma, Alzheimer's disease, Parkinson's disease, depression, schizophrenia, autism, and cancer.

Asthma causes swollen and inflamed vocal folds that do not vibrate appropriately during voice generation. The voice sound becomes hoarse and impaired. A detailed investigation of this issue can be found in [24]. That work analyzes speech segments for asthma patients of variable lengths. The speech segments include five minutes conversation, a monologue, and counting numbers. Voice parameters, namely onset time, word duration, pause time, and total activity duration for normal and asthmatic subjects, are considered in that work. The results show that asthmatic subjects demonstrate longer pauses between speech segments, produce fewer syllables per breath, and spend more time in voiceless ventilator activity than their healthy counterparts.

Another major cause of voice disability is Alzheimer's disease [25]. The common symptoms of Alzheimer's disease are memory loss, confusion, inability to retain information, aggressiveness, trouble with language, and mood swings. Studies show that

Alzheimer's disease also causes aphasia [26]-[27]. Although memory impairment has generally been considered the major symptom of Alzheimer's disease, it is now reported that language deficits occur in about 8%-10% of Alzheimer's patients. Hence, they can be used as a primary symptom to detect this disease at its early stages [28]-[30]. Similar work shows that about 5% of Alzheimer's patients' language capacity is steadily impaired during the developing period of this disease [31]. Other works [32]-[33] also show that disrupted language is an early symptom of Alzheimer's disease. A comprehensive study on voice disability due to Alzheimer's disease can be found in [34].

Parkinson's disease is another major cause of voice disability. Generally, Parkinson's disease causes the loss of neurons in the brain, affecting the motor and non-motor body functions of the human body. Parkinson's patients face problems related to recognition, behavioral changes, insomnia, and sensory difficulties [35]. These symptoms are often followed by other symptoms, including slower movement, rigidity, tremor, and postural instability. This disease also affects patients' muscles in the voice generation system; hence, patients speak slowly, loosely, and breathily. The patients even find difficulty in pronouncing words correctly. They also generate undesired voices due to their faulty vocal folds [36]-[38]. Recent research shows that voice disability can indicate an early symptom of Parkinson's disease [39].

Depression is a psychiatric disorder that affects a human's mood, behavior, thoughts, senses, ailments, and feelings. This disease can make a patient anxious, fatigued, irritable, and worried. The patient may have a problem with decision-making, memorizing, and losing interest in activities. Studies show that depression can also affect the patients' voice system [40]. The patients speak softly, slowly, hesitatingly, and monotonously. They often stutter and mute in the middle of a sentence [41]. Hence, voice features, including pitch, energy, speaking rate, formants, and power spectral density, can be used to identify a depressed patient [42]-[43]. It is also shown in [44] that acoustic patterns of voice for depressed patients can be used to track the disease from an early stage to a treatment stage. These findings suggest that acoustic measures of patients' voices can provide an objective procedure to evaluate depression.

Schizophrenia is a neurodevelopmental disorder that affects voice disability [45]. Schizophrenia patients usually suffer from delusions, hallucinations, movement disorder,

and disorganized speech. They even sometimes talk about strange and unusual ideas. A study [46] shows that speech fluctuations can be biomarkers for schizophrenia. Hence, advanced signal processing techniques and artificial intelligence can be employed to investigate voice features that contain substantial emotional information about a schizophrenia patient. In [47], two spectral features, MFCCs and Linear Predictive Coding (LPC), have been used to separate patient groups from the normal group. It is shown that MFCC scores are significantly lower, and LPC scores are considerably higher in the schizophrenic patient group than in the normal group.

Autism spectrum disorder is another neurodevelopmental disorder that can affect voice disability. One of the earliest works on autism can be found in [48]. In that work, autism is characterized by impairment in social interaction, behaviors, and communication skills. Autistic patients often say something irrelevant that does not match the situation [49]. Hence, speech and prosody-voice profiles can characterize autistic patients and patients with Asperger Syndrome (AS) [50]. It is shown in the work that the patients suffering from these two diseases cause residual articulation distortion errors, non-understandable utterances, and inappropriate phrasing, stress, and resonance. Another study [51] correlates acoustic measurements to communication impairment due to autism. That work shows that fundamental frequency variation in the narrative of the autistic patient can be related to the intelligence quotient (IQ) and verbal abilities of autistic patients. In that work, PRAAT [52] software was used to extract the fundamental frequency of several autistic and controlled patients. The comparison shows that the fundamental frequency of autistic patients has a higher standard deviation than controlled patients. Inflection of voice, the pattern of pauses, relative duration of syllables, relative loudness, and rhythm are often included in the prosodic features of voice [53]. Hence, prosody, particularly prominence, and prosodic contours, can be used to investigate the communicative intent and conversational skills of autistic patients. The results in [54] show that abnormal prosody is the core deficit in autistic patients.

Cancer is another cause of voice disability. Studies show that voice features can relate to the cancer stage [55]. Based on the speech content analysis of 71 patients, it is demonstrated that voice features can be used to detect signs of cancer in the head and neck. The results show that the systematic quantification of lexical choice can be used as an

indicator for cancer detection. Automatic speech recognition has also been used in [56] to detect neck and head cancer. The authors conclude that speech recognition can provide the percentage of correctly recognized words of a sequence. The same work shows that the cancer patients have significantly lower word recognition rates than the control group. Hence, automatic speech recognition can serve as a reasonable means to objectify and quantify cancer patients. Another study [57] shows that the role of emotional expression and cancer progression are related. The study used the voice samples of 25 breast cancer survivors and 25 controlled patients. The results show that cancer patients use fewer inhibition words than controlled patients. The results also show that cancer diagnosis and treatment can alter the emotionally expressive behavior of a patient.

2.2 Current Procedures for Voice Pathology Detection

To detect voice disability, physicians use common procedures, namely, laryngoscopy, laryngeal electromyography, stroboscopy, and imaging tests [58]. In laryngoscopy, the patient's throat is examined by a light source. There are three types of laryngoscopy: direct laryngoscopy, indirect laryngoscopy, and fiber optic laryngoscopy. A direct rigid laryngoscopy procedure is used to examine patients' vocal cords or larynx. A laryngoscope is a rigid and hollow tube with a light attached at the top. Using this tool, the physicians can examine behind the patient's tongue and down the throat to the vocal cords, as shown in Fig. 2.1. With indirect laryngoscopy, a small mirror is held at the back of the throat illuminated by a light source. With fiber optic laryngoscopy, a laryngoscope is inserted through the nose down into the throat, as shown in Fig. 2.2.

Using laryngeal electromyography (LEMG), electrical activity in the throat muscles is measured. LEMG is a useful diagnostic tool for examining the human larynx. The larynx is a complex system consisting of various muscles that help humans to speak. Even a minor absence of vocal cord movement can cause respiratory and vocal problems. LEMG can help to find the original cause of reduced muscle movement. Major reasons for reduced vocal fold movement are related to disruption of the laryngeal and superior laryngeal nerve. By using LEMG, it is possible to determine the vocal fold's tonicity. In this method, a thin needle is pierced into the neck muscles, and the conductivity of the muscles is measured with electrodes.

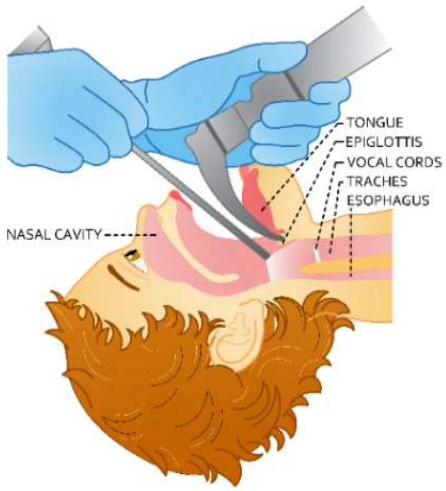


Figure 2.1 The direct laryngoscopy.

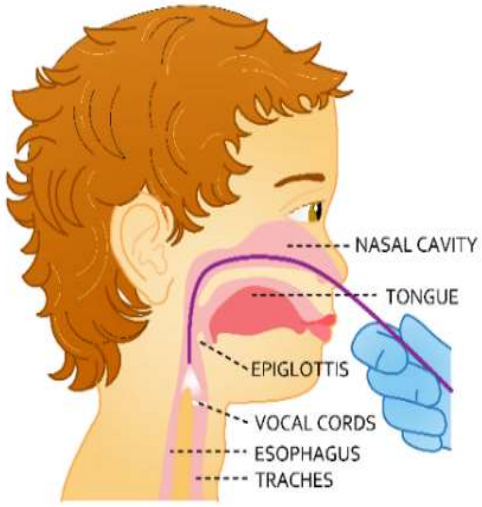


Figure 2.2 The Fiber optic laryngoscopy.

With stroboscopy, a light source and a video camera are used to examine the vocal cord vibration. The vocal folds vibrate very fast during voice production, which is impossible to notice with the naked eye. Hence, a stroboscopy is used. A bright flashing light is used during this procedure to illuminate the vocal folds. By taking multiple snapshots at different phases of the vibration, it is possible to examine the movement of the vocal folds. Medical imaging techniques, including X-rays, computerized tomography

(CT) scans, and magnetic resonance imaging (MRI), are also used to diagnose voice disability. These medical imaging techniques are very effective in examining the growths of tissue or other problems in the throat.

2.3 Experimental Setup of Common Voice Features for Voice Pathology Detection

To detect voice disability, researchers use several voice features. The most common voice features are MFCCs, spectrogram, formants, wavelets, LPC, perceptual linear prediction (PLP), relative spectral transform – PLP (RASTA-PLP), Jitter, Shimmer, glottal-to-noise ratio (GNR), HNR, cepstral based HNR (CHNR), noise energy to total energy ratio (NNR), ZCR, linear frequency cepstral coefficient (LFCC), and Teager energy operator (TEO). These voice features are briefly described in the following subsection. To explain these voice features, two voice samples have been used – one for a pathological baby and the other for a normal baby. These voice samples are shown in Fig. 2.3. The two babies, in the age group of 6-8 years, are asked to narrate the same story. The samples are taken, for feature extraction, from the beginning of their story narration.

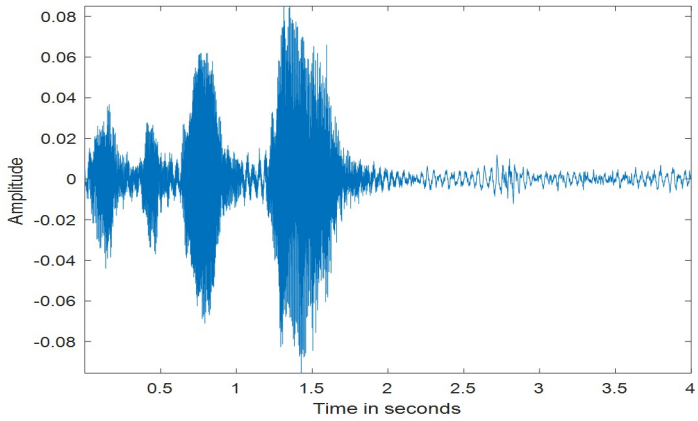
2.3.1 The Mel-frequency Cepstral Coefficients (MFCCs)

The MFCCs have been widely used in voice disability detection algorithms. The main advantage of MFCCs over other voice features is that they can completely characterize the shape of the vocal tract. Once the vocal tract is accurately characterized, one can estimate an accurate representation of the phoneme being produced by the vocal tract. The shape of the vocal tract manifests itself in the envelope of a short-time power spectrum, and the MFCCs accurately represent this envelope.

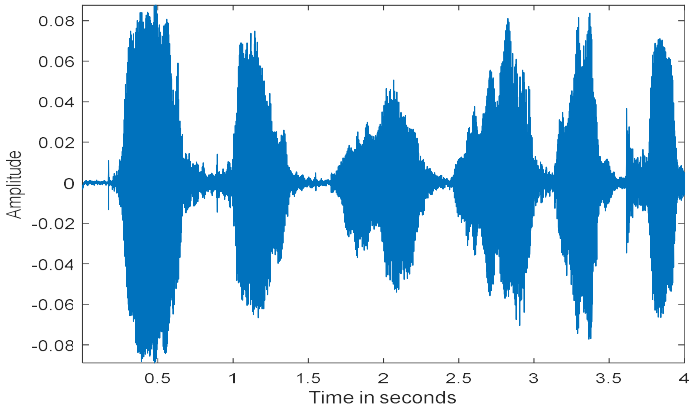
The MFCCs are determined by the following procedure [2]. The voice sample $x[n]$ is first windowed with an analysis window $w[n]$ and the STFT, $X(n, \omega_k)$ is computed by

$$X(n, \omega_k) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega_k n}, \tag{2.1}$$

where $\omega_k = \frac{2\pi k}{N}$ with N is the discrete Fourier transform (DFT) length. The magnitude of



(a) Normal sample



(b) Pathological sample

Figure 2.3 The voice samples used in the analysis.

$X(n, \omega_k)$ is then weighted by a series of filter frequency responses whose center frequencies and bandwidth are roughly matched with those of auditory critical band filters called Mel-scale filters. The next step is to compute the energy in STFT weighted by each Mel-scale filter frequency response. The energy for each speech frame at time n and the lth Mel-scale filter is given by

$$E_{mel}(n, l) = \frac{1}{A_l}\sum_{k=L_l}^{U_l}|V_l(\omega_k)X(n, \omega_k)|^2, \tag{2.2}$$

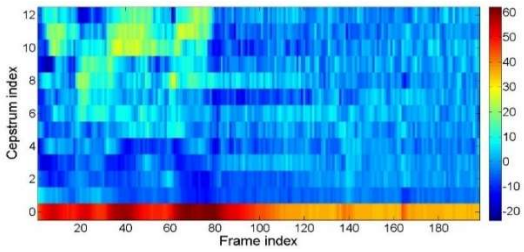
where $V_l(\omega)$ is the frequency response of the lth Mel-scale filter, L_l and U_l are the lower and upper-frequency indices over which each filter is nonzero, while A_l is defined as

$$A_l = \sum_{L_l}^{U_l} |V_l(\omega_k)|^2 \tag{2.3}$$

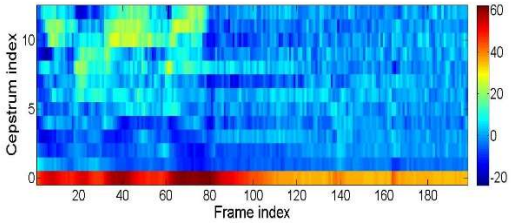
The cepstrum, associated with $E_{Mel}(n, l)$, is then computed for the speech frame at time n by

$$C_{mel[n,m]} = \frac{1}{R} \sum_{l=0}^{R-1} log\big(E_{Mel}(n, l)\big) cos \frac{2\pi ml}{R}, \tag{2.4}$$

where R is the number of filters. An example of MFCCs of normal voice and pathological voice samples (presented in Fig. 2.3) are shown in Fig. 2.4. The plot shows the distribution of the magnitudes for MFCCs with frame index and cepstrum index. It shows that the magnitude of the MFCCs is high with the lower frame indices for normal voice. On the other hand, MFCCs for pathological voice are randomly distributed among a wide range of frame indices. Hence, MFCCs are extensively used in several works for discriminating pathological voices from normal voices.



(a) The MFCCs of normal voice



(b) The MFCCs of pathological voice

Figure 2.4 The MFCCs of normal and pathological voice samples.

2.3.2 The Spectrogram

A speech waveform consists of a sequence of different events that vary with time. This time-varying nature corresponds to highly fluctuating spectral characteristics over time.

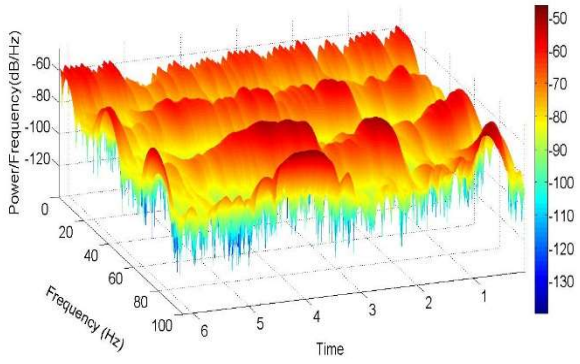
Hence, a single Fourier transform cannot capture this type of fast time-varying signal, and STFT is used instead [59]. The STFT consists of a separate Fourier transform for pieces of the waveform under the sliding window. Then, the spectrogram of the voice signal is derived from STFT by

$$S(\omega) = |X(m, \omega_k)|^2 \tag{2.5}$$

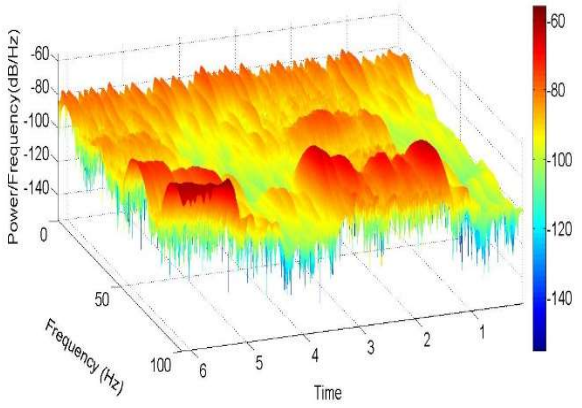
The spectrogram can be presented in a 3-D plot to show the distribution of power densities with respect to time and frequency, as shown in Fig. 2.5. It is observed in the figure that the power density distribution of the voice signal widely varies with time and frequency, and it can be used to distinguish between normal and pathological voices. It is also seen in the figure that power distribution for normal voices is more uniform with respect to time and frequency. Hence, a spectrogram is an excellent indicator for discriminating pathological voices from normal voices.

2.3.3 The Formants

The formant frequency or simply formant analysis is another vital voice feature researchers have used for voice pathology detection. The formant frequencies are the resonance frequencies of the vocal tract that change with different vocal tract configurations [60]. The formant usually refers to the entire spectral contribution of a resonance. The peaks of the spectrum for vocal tract response correspond approximately to its formants. The formants can be plotted with respect to frequency, as shown in Fig. 2.6. The formant plot shows distinct peaks at specific frequencies. It also shows that the peaks are separated by some frequency band and are of decreasing magnitudes. The formant plot shows that the pathological voice exhibits very distinct formants compared to the normal voice. For example, the first three peaks are closely located and are of almost the same magnitude for pathological voice. On the other hand, normal voice shows peaks located at nearly equal distances and with decreasing magnitudes.



(a) Spectrogram of normal voice sample



(b) Spectrogram of pathological voice sample

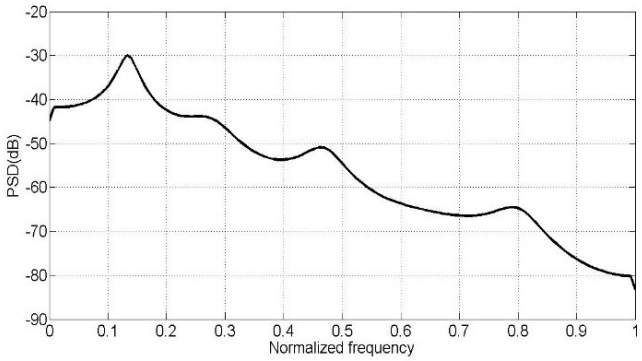
Figure 2.5 The Spectrograms of normal voice and pathological voice.

2.3.4 The Wavelet Analysis

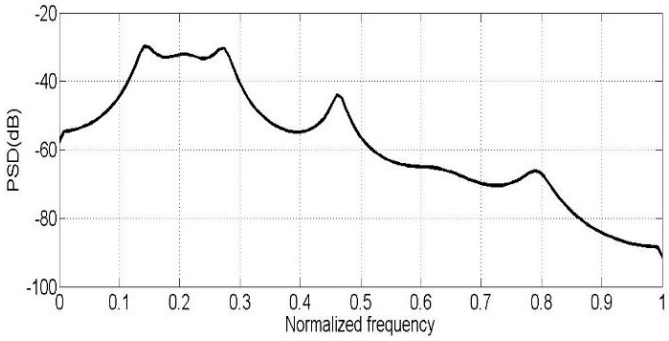
The wavelet transform is another essential tool used in voice disability detection. Its main advantage over the Fourier transform is that the wavelet can provide accurate information about the fast fluctuations of signals in the time domain. It maps a time function into two functions: scale, a, and translation, b [61]. The CWT of a signal $f(t)$ is defined as

$$W(a,b) = \int_{-\infty}^{\infty} f(t)\varphi_{ab}(t)d\varphi \ , \tag{2.6}$$

where $W(a,b)$ is the wavelet transform and $\varphi_{ab}(t)$ is the mother wavelet, which is defined as



(a) Formants of normal voice

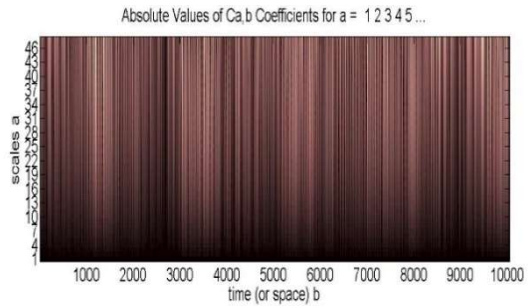


(b) Formants of pathological voice

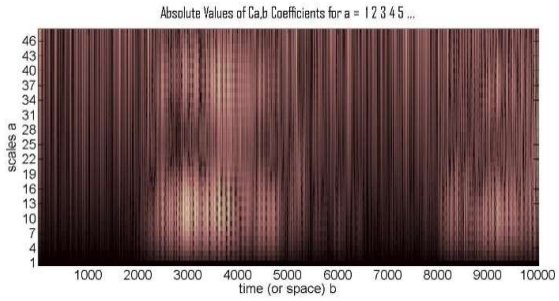
Figure 2.6 The comparison of the formants.

$$\varphi_{ab}(t) = \frac{1}{\sqrt{a}}\varphi\left(\frac{t-b}{a}\right) \tag{2.7}$$

A scaled version of the function $\varphi(t)$ with a scale factor of a is defined as $\varphi\left(\frac{t}{a}\right)$. A wavelet is a valuable tool for investigating the discontinuity in pathological voice. The plot of wavelet coefficients for normal and pathological voices is shown in Fig. 2.7. The discontinuity in the pathological voice is more visible in the plots. Fig. 2.7(b) shows some discontinuity in voice signals in the range of 2500-5000 samples and 8000-8500 samples.



(a) Wavelet coefficients of normal voice.



(b) Wavelet coefficients of pathological voice.

Figure 2.7 Wavelet analysis comparisons.

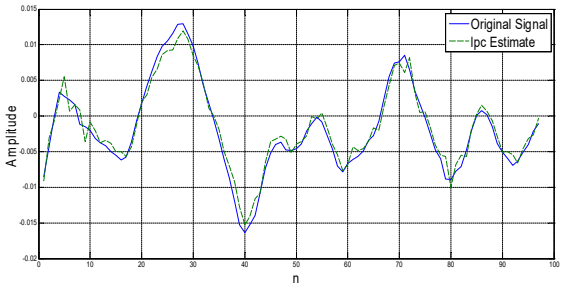
2.3.5 The Linear Predictive Coding (LPC)

Primarily, LPC has been introduced to compress digital signals for efficient transmission and storage. However, LPC has also become one of the most powerful speech analysis techniques and gained popularity as a formant estimator [62]. The LPC method is based on modeling the vocal tract as a linear all-pole infinite impulse response (IIR) filter, which is defined by

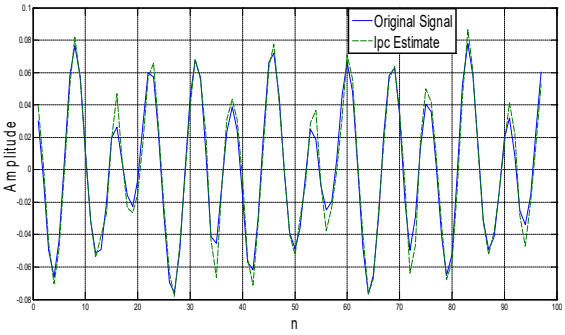
$$H(z) = \frac{G}{1+\sum_{k=1}^{p} a_p(k) z^{-k}}, \tag{2.8}$$

where p is the number of poles, G is the filter gain, and $a_p(k)$ are the coefficients. Given a short-time segment of a speech signal (i.e., 20 ms) sampled at a 8 kHz sampling rate, a speech encoder determines proper excitation function, pitch period for voiced speech, gain parameter G, and the coefficients $a_p(k)$. The LPC is computed using the least mean-squared error approach [63]. This approach approximates the speech signal as a linear

combination of its previous samples. LPC plots are generated by PRAAT [52] software, and the plots (original signal and estimated signal) are shown in Fig. 2.8. This figure shows that LPC coefficients have distinctively varying magnitudes in some portions of voice signal. However, the magnitude is not significant for other parts of the voice signal. The magnitude distribution can be used to differentiate pathological voice from normal voice.



(a) Normal sample



(b) Pathological sample

Figure 2.8 The LPC coefficients.

2.3.6 The Perceptual Linear Prediction (PLP)

PLP, introduced by Hermansky [64], models human speech based on the concept of the psychophysics of hearing. The primary function of PLP is to discard irrelevant information contained in the speech. PLP has spectral characteristics that are transformed to match the human auditory system, unlike LPC. Hence, PLP is more adapted to human hearing compared to LPC. The other main difference between PLP and LPC is that both use two different types of transfer functions. For example, the LPC model assumes an all-pole transfer function of the vocal tract with a specified number of resonances within the analysis band. On the other hand, the transfer function of PLP is also an all-pole model;

however, it approximates power distribution of equal magnitude at all frequencies of the analysis band. The detailed steps of PLP computation are shown in Fig. 2.9.

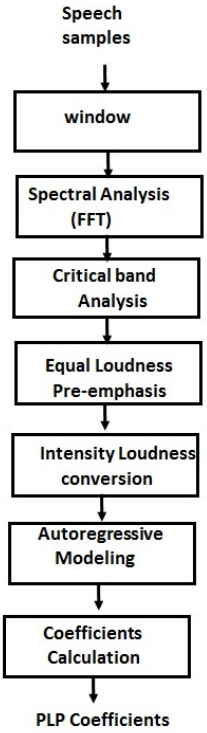


Figure 2.9 The computation of PLP.

To compute the PLP, the speech samples are weighted by a window function and transformed into the frequency domain using Fast Fourier Transform (FFT). Then the power spectrum is determined by

$$P(\omega) = \left[Re\left(S(\omega)\right)\right]^2 + \left[Im\left(S(\omega)\right)\right]^2, \tag{2.9}$$

where $S(\omega)$ is the Fourier transform of the windowed voice signal. A frequency warping into the Bark scale is applied. The first step is a conversion from frequency to bark scale frequency, which is a better representation of the human hearing resolution in frequency. The bark frequency, $R(\omega)$ [64] corresponding to an audio frequency, is given by

$$R(\omega) = 6 \ln\left[\frac{\omega}{1200\pi} + \sqrt{\left(\frac{\omega}{1200\pi}\right)^2 + 1} \quad\right]. \tag{2.10}$$

The warped auditory spectrum is then convoluted with the power spectrum of the simulated critical-band masking curve to simulate the critical-band integration of human hearing. The smoothed spectrum is then downsampled. The three steps: frequency warping, smoothing, and sampling, are usually integrated into a single filter bank. An equal-loudness pre-emphasis is applied to the filter-bank outputs. The equalized values are then warped and processed by a linear predictor (LP). Finally, the cepstral coefficients are obtained from the LP coefficients by a recursive method.

2.3.7 The Rasta Perceptual Linear Prediction (RASTA-PLP)

Another popular speech feature used in voice disability detection is known as RASTA-PLP. A special bandpass filter, called the RASTA filter, is used in computing the RASTA-PLP. An example of the system function for the RASTA filter is defined by

$$H(z) = 0.1z^4 * \frac{2+z^{-1}-z^{-3}-2z^{-4}}{1-0.98z^{-1}} \tag{2.11}$$

The lower cut-off frequency of the filter determines the fastest spectral change ignored at the output. On the other hand, the higher cut-off frequency determines the fastest spectral change preserved in the output. The filter's primary function is to suppress the frequency that varies more quickly or slowly in the voice signal. The steps of computing the RASTA-PLP are shown in Fig. 2.10.

The RASTA-PLP is computed in the following steps: (a) compute the critical-band power spectrum, (b) transform spectral amplitude through a compressing static nonlinear transformation, (c) filter the time trajectory for each transformed spectral component, (d) transform the filtered speech representation through expanding static nonlinear transformation, (e) multiply by equal loudness curve and raise to the power 0.33 to simulate the power of law for hearing, (f) compute all-pole model of the resulting spectrum, following the conventional PLP technique. The plots of RASTA-PLP for normal and pathological voices are shown in Fig. 2.11. The RASTA-PLP plots for normal and pathological voice samples show apparent differences in magnitude for time and frequency.

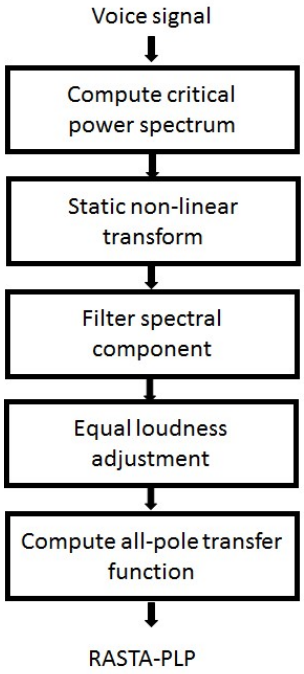
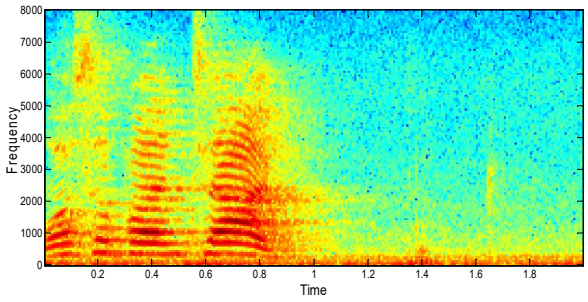
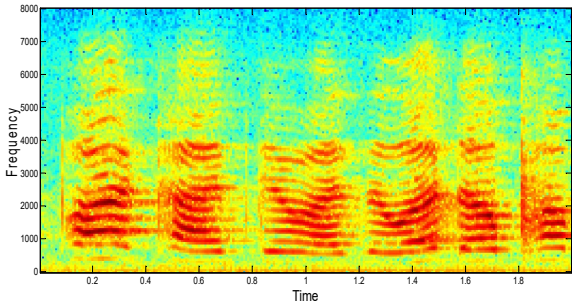


Figure 2.10 The computation of RASTA-PLP [65].



(a) Normal



(b) Pathological.

Figure 2.11 The RASTA-PLP spectra comparison.

2.3.8 The Jitter

Jitter reflects the variation of successive periods in the voice signal. Determining Jitter needs to detect the timing of the fundamental period. After the determination of onset time for the glottal pulses, Jitter can be determined for its several measured shapes given by the expressions shown below.

Jitter (local, absolute): It is defined by (2.12), and it represents the average absolute difference (over N periods) between two consecutive periods (i.e., $T_i - T_{i-1}$). The T_i is extracted from the period length, F_0 and N is the number of extracted periods. This is also known as *Jitta*. This parameter can be used to detect voice pathology by comparing it with a threshold value. The threshold value to detect pathologies in adults is 83.2 μs, as reported in [66]-[67].

$$\text{Jitta} = \frac{1}{\text{N}-1}\sum_{\text{i}=1}^{\text{N}-1}|\text{T}_\text{i} - \text{T}_{\text{i}-1}| \tag{2.12}$$

Jitter (local): It represents the average absolute difference between two consecutive periods divided by the average period. It is also known as *Jitt*, given by (2.13), and has 1.04% as the threshold limit for detecting pathologies.

$$\text{jitt} = \frac{\text{jitta}}{\frac{1}{\text{N}}\sum_{\text{i}=1}^{\text{N}}\text{T}_\text{i}}, \tag{2.13}$$

where T_i is the duration in seconds for each period.

Jitter (*rap*): It represents the average absolute difference of one period and the average of periods with its two neighbors divided by the average period. The *rap* is defined by (2.14) and its threshold value to detect pathologies is 0.68%.

$$\text{rap} = \frac{\frac{1}{\text{N}-1}\sum_{\text{i}=1}^{\text{N}-1}\left|\text{T}_\text{i} - \left(\frac{1}{3}\sum_{\text{n}=\text{i}-1}^{\text{i}+1}\text{T}_\text{n}\right)\right|}{\frac{1}{\text{N}}\sum_{\text{i}=1}^{\text{N}}\text{T}_\text{i}} \tag{2.14}$$

Jitter (*ppq5*): The *ppq5* is defined by (2.15), and it represents the average absolute difference between a period and the average containing its four nearest neighbor periods divided by the average period.

$$\text{ppq5} = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}\left|T_i - \left(\frac{1}{5}\sum_{n=i-2}^{i+2} T_n\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \tag{2.15}$$

2.3.9 The Shimmer

Shimmer is another voice feature widely used in voice disability detection [66]-[67]. Unlike Jitter, Shimmer focuses on the peak values of a signal. To determine Shimmer parameters, the algorithm begins by defining the onset time of a signal's glottal pulses and the signal's respective magnitude at that sample. Then, the algorithm is applied to determine the values of each parameter of Shimmer. Several Shimmer parameters are defined as follows.

Shimmer (local): It represents the average absolute difference between the amplitudes A_i and A_{i+1} of two consecutive periods T_i and T_{i+1}, divided by the average amplitude. It is called a 'Shim', and this parameter is set to 3.81% as the limit for detecting pathologies. The expression of *Shim* is given by

$$\text{Shim} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{2.16}$$

Shimmer (local, dB): It represents the average absolute difference of base 10 logarithms for the difference between two consecutive periods and is called *ShdB*. The limit to detect pathologies is 0.350 dB. The *ShdB* (local dB) is given by

$$\text{ShdB} = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20 * \log\left(\frac{A_{i+1}}{A_i}\right)\right| \tag{2.17}$$

Shimmer (*apq3*): It represents the quotient of amplitude disturbance within three periods. In other words, the average absolute difference between a period's amplitude and its two neighbors' mean amplitudes is divided by the average amplitude. It is given by

$$\text{apq3} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A_i - \left(\frac{1}{3}\sum_{n=j-1}^{i+1} A_n\right)\right| \times 100}{\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{N} A_i} \tag{2.18}$$

Shimmer (*apq5*): It represents the ratio of perturbation amplitude of five periods. In other words, the average absolute difference between the amplitude of a period and the

mean amplitudes of it and its four nearest neighbors is divided by the average amplitude. The *apq5* is given by

$$\text{apq5} = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}\left|A_i-\left(\frac{1}{5}\sum_{n=i-2}^{i+2}A_n\right)\right|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \tag{2.19}$$

2.3.10 The NNE, GNR, HNR, and CHNR

NNE is the ratio between the noise energy and the signal's total energy measured in dB) [68]. The noise energy (between two harmonics) is directly obtained from the spectrum. Within a harmonic, the noise energy is assumed to be the mean value of both adjacent minima in the spectrum. If the harmonics are broadened because of Jitter or Shimmer, the energy outside the window defined for the harmonic is erroneously assigned to noise energy. Hence, the noise measured by NNE appears to be increasing. To overcome this problem, it is common practice to vary the frequency range to obtain the best discrimination between normal and pathological (glottal cancer) voices.

The HNR is determined based on the mathematical fundamentals presented by Boersma [66]. It is determined by the detection of the autocorrelation function for the voice signal. The HNR is defined by

$$\text{HNR} = 10.\log\frac{AC_V(T)}{1-AC_V(T)}, \tag{2.20}$$

where the $AC_V(T)$ is the peak at the index position corresponding to the period of the signal.

Roughly speaking, CHNR is the cepstrum-based HNR [69] and is the inverse of NNE. It is the ratio between total energy and energy of noise (both measured in dB). However, the energies are obtained differently. At first, the cepstral peaks at the fundamental period, and its multiples are removed. Essentially, the spectral energy between harmonics below the lines that connect minima is considered noise energy. Therefore, the inverse CHNR is generally larger than NNE. Due to Jitter and Shimmer, the harmonics are broadened, and the minima of the spectrum are less deep. Hence, in the presence of Jitter and Shimmer, the noise energy is overestimated by CHNR. It is based on the correlation coefficient for Hilbert envelopes of different frequency bands. The parameter indicates whether a given voice signal originates from the vibrations of vocal folds or turbulent noise

generated in the vocal tract and is thus related to breathiness. Therefore, it is called the GNE. The GNE factor is calculated in the following way (a) down-sampling speech signal to 10 kHz, (b) inverse filtering of the speech signal, (c) calculating the Hilbert envelopes, (d) calculating the cross-correlation function between such envelopes, (e) picking the maximum of each correlation function, and (f) picking the maximum from the maxima in step.

2.3.11 The Zero Crossing Rate (ZCR)

In the context of discrete-time signals, a zero crossing is said to occur when successive samples have different algebraic signs. The rate of zero-crossing is a simple measure of the frequency content of a signal. The zero-crossing rate is a measure of the number of times in each time interval divided by the frame that the amplitude of speech signals passes through a value of zero [70]. The zero-crossing rate is defined by

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]| w(n-m), \tag{2.21}$$

where $\text{sgn}[x(n)] = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases}$, and $w(n) = \begin{cases} \frac{1}{2N}, & 0 \leq n \leq N-1 \\ 0, & \text{otherwise} \end{cases}$.

Based on the speech production model, we conclude that the energy of voiced speech is concentrated below 3 kHz because of the spectrum fall introduced by the glottal wave. On the other hand, unvoiced speech is concentrated in the higher frequencies. Since high frequencies imply high zero crossing rates and low frequencies imply low zero-crossing rates, there is a strong correlation between zero-crossing rates and the energy distribution of a signal with respect to frequency.

2.3.12 The Linear Frequency Cepstral Coefficients (LFCCs)

LFCC is computed as MFCC with a filter bank of 40 bands MFCC-FB40 [71]. The only difference is that the Mel Frequency warping step is skipped [72]. This algorithm implements the desired frequency range by a filter bank of 40 equal-width and equal-height linearly spaced filters. The bandwidth of each filter is 164 Hz, and the whole filter bank covers the frequency range of 133-6857 Hz. The LFCC is computed by using the following steps: (a) apply N-point DFT to the discrete-time domain input signal $x(n)$, (b) apply triangular filtering, (c) compute logarithmically compressed filter bank outputs, and (d)

apply discrete cosine transform (DCT) to the filter bank outputs to obtain LFCC FB-40 parameters.

2.3.13 The Teager Energy Operator (TEO)

TEO is introduced by Teager. Detailed information on TEO and its development can be found in [73]-[74]. In the discrete-time domain, the TEO is given by

$$\varphi[x(n)] = x^2(n) - x(n+1)x(n-1) \tag{2.22}$$

There are several applications of the TEO, including tracking information content in speech signals. This operator can track vowels and formants in the voice signal. It is also used to find the center frequency and bandwidth of the formants. Recently, the TEO has been used in implementing a voice pathology detection algorithm.

2.4 The Classifiers

An essential final purpose of voice signal analysis is to classify a given signal into one of a few known categories and to arrive at a diagnostic decision about voice disability. The classification of a given voice signal into one of many categories is beneficial in diagnostic procedures. Pattern recognition or classification algorithms are used for this purpose. Several classifiers have been used in voice disability detection. The commonly used classifiers are explained in this section.

2.4.1 Support Vector Machine

SVM applies the statistical concept of support vector to classify data [75]. It uses the concept of the supervised learning model. The supervised learning algorithm maps between input and output by using a function. SVM uses a training algorithm to build a model based on the provided data. Once the learning model is established, SVM can classify data into two categories. Generally, SVM constructs a hyperplane for decision-making. This type of decision-making is called classification. The hyperplane can be a linear line or a non-linear line. Intuitively, the performance of SVM depends on the separation defined by the optimum hyperplane. Hence, a good separation provides high accuracy. A good separation is defined as the largest distance to the nearest training-data point of any class. The larger margin between two data sets minimizes the error produced by the classifier.

2.4.2 Gaussian Mixture Model and GMM-Universal Background Model

GMM is a probabilistic model for classifying normally distributed data within an overall data [76]. It is a widely used algorithm to classify voice features. Unlike SVM, the GMM is an unsupervised algorithm. An unsupervised algorithm does not need prior knowledge about the subpopulation of data. The model learns the subpopulation automatically. Generally, the GMM algorithm is considered suitable for modeling extensive real-world data. Notably, this algorithm is suitable for datasets that are Gaussian distributed. The GMM algorithm exploits the theoretical and computational benefits of Gaussian models.

GMM-UBM framework is a modified version of the GMM model. GMM-UBM can handle a large dataset; hence, it is considered suitable for classifying large voice samples extracted from many speakers [77]. Once voice features are extracted, speaker-specific models are then adapted from UBM using maximum a posterior probability algorithm (MAP). This MAP algorithm has mainly two steps. In the first step, information about the parameters is estimated. In the second step, the new information regarding the parameters is mixed with the old parameters, and the model is updated. This kind of mixing is highly influenced by language-specific data.

2.4.3 Artificial Neural Network (ANN)

In many practical applications, no prior probabilities of patterns belonging to a certain class are available. Hence, no general classification rule can be used for pattern recognition. In such applications, conventional pattern classification methods are not well suited. However, ANN is considered an effective tool for solving such classification problems [78]. ANN possesses some properties, including experience-based learning and fault tolerance. These properties make ANN particularly suitable for solving classification problems.

ANN has one hidden layer and one output layer for pattern classification. Based on an initially provided training dataset, the network learns similarities among patterns directly from their instances. Classification rules are determined from training data without prior knowledge of patterns in the data. ANN is trained by an algorithm called backpropagation. Backpropagation is a method used in artificial neural networks to calculate weights that are used in the network. The backpropagation is also known as

backward propagation of errors because the error is computed at the network's output and distributed backward through the upper layers.

2.4.4 Hidden Markov Model (HMM)

HMM is a statistical model used to model data that can be defined by the Markov process with unobserved states, called hidden states. It can be represented by a dynamic Bayesian network. The mathematical formulation of HMM can be found in [79]-[80]. The differences between the simple Markov model and HMM are as follows. The states of simple Markov models are directly visible to an observer; therefore, these models only consider state transition probabilities. On the other hand, the states of HMM are not directly visible. However, the output of HMM is in the form of data. Each state has a probability distribution over the possible output data. Therefore, an HMM generates some sequence of data containing the series of states. Some typical applications of HMM models include speech recognition, handwriting recognition, and gesture recognition. Recently, they are also being used in voice disability detection algorithms.

2.4.5 Deep Neural Network (DNN)

DNN is an ANN with multiple layers [81]-[82]. It can find both linear and nonlinear relations between input and output data. DNN is trained through different layers to find the probability of each output. In DNN, each mathematical relation is considered a layer. A complex DNN uses many layers to model a complex non-linear relationship between input and output. The architecture of DNN generates compositional models based on the data. The extra layers used by DNN enable the composition of features from lower layers. DNN is typically a feedforward network, where data flows from the input layer to the output layer without a feedback loop. At first, DNN creates a map of virtual neurons, assigns random weights, and then establishes the connection between them. The weights and inputs are multiplied, and the output is returned. If the network fails to recognize a particular pattern, the algorithm adjusts weights, and the whole process repeats.

2.4.6 Convolutional Neural Network (CNN)

CNNs are deep artificial neural networks. CNNs are commonly used to classify data, cluster them by similarity, and perform object recognition [83]. Some applications of CNNs include identifying faces, individuals, street signs, tumors, and platypuses. CNNs

are popularly applied in voice analysis and image recognition. It is particularly suitable for spectrogram analysis of voice signals.

2.4.7 Probabilistic Neural Network (PNN)

PNN is designed to solve classification problems using a statistical memory-based approach. It can use both supervised and unsupervised algorithms [84]. In PNN, a Parzen window is used to determine a parent probability distribution function (PDF) for each population class. Then, Bayes' rule is employed to allocate the class with the highest posterior probability to new input data. This is done to minimize the probability of misclassification. PNN uses the Kernel functions that make it suitable for discriminant analysis and pattern recognition. Hence, it is popularly used in voice disability detection algorithms.

With the given input, the first layer of PNN computes the distance from the input vector to the training input vectors. This produces a vector to indicate the proximity of input to training input. The second layer sums the contribution for each class of inputs and produces net output as a vector of probabilities. Finally, a complete transfer function on the output of the second layer picks the maximum of these probabilities and produces output for non-targeted classes. There are several advantages of PNN over perceptron networks. PNN is faster than other multilayer perceptron networks. It is also more accurate than multilayer perceptron networks.

2.4.8 Deep Belief Network (DBN)

In machine learning, DBN is a multilayer deep neural network with a connection between layers [85]. When trained on a set of examples without supervision, DBN can reconstruct its inputs based on probabilistic models. After this learning step, DBN can be further trained with supervision to perform classification. A DBN can be viewed as a composition of simple and unsupervised networks based on the concept of restricted Boltzmann machines (RBMs). In RBMs, each sub-network's hidden layer serves as a visible layer for the next layer. RBMs consist of a visible input layer connected to a hidden layer with connections in between. This type of architecture leads to a fast unsupervised training procedure. The contrastive divergence is applied to each sub-network in turn, starting from

the lowest pair of layers. DBNs can be trained greedily and hence, are considered effective deep learning algorithms.

2.4.9 Generalized Regression Neural Network (GRNN)

GRNN [86] is a memory-based network to estimate continuous variables and converges to an underlying regression surface. GRNN is a one-pass learning algorithm with a parallel structure. GRNN algorithm provides a smooth transition of data from one state to another state, even in multidimensional space. The algorithm uses linear and nonlinear regression models to predict, map, and interpolate the model. The structure of GRNN is similar to that of PNN. The main difference is that PNN determines decision boundaries between patterns, whereas GRNN estimates values for continuous variables. GRNN has the following several advantages over other neural networks. The network learns in one pass through the data and converges to a conditional mean regression surface as more examples are learned. The estimate is bounded by a minimum and a maximum number of observations. The estimate cannot converge to a poor solution corresponding to a local minimum of the error criteria. The main disadvantage of the GRNN algorithm is that it requires substantial computation to evaluate the algorithms.

2.4.10 Bayesian Classifier

The Bayesian classifier is another popular classifier used to classify data based on common features [87]-[88]. The Bayesian classifier is a probabilistic model, where the classification is a latent variable related to the observed variables by a probabilistic model. The Bayesian classifier works based on the following principles. If an agent knows the class, it can predict the values of other features. If it does not know the class, a rule called Bayes' rule is applied to predict the class. In a Bayesian classifier, the learning agent builds a probabilistic model based on the provided data features and uses the model to predict the classification of a new dataset. Then, classification becomes an inference in the probabilistic model. A naive Bayesian classifier is based on the assumption that input features are conditionally independent of each other. It is a belief network where the features are the nodes, the target variable has no parents, and the classification is the only parent of each input feature.

2.4.11 The K-Means Clustering

The k-means clustering is a method of vector quantization that is popularly used for cluster analysis in data mining [89]. The k-means clustering aims to partition n observations into k clusters. Each observation belongs to a cluster with the nearest mean. This results in partitioning a data space into cells called Voronoi cells. The k-means clustering algorithms are computationally expensive. However, some efficient heuristic algorithms have been proposed to reduce the computations by converging quickly to a local optimum.

2.4.12 The Decision Tree Algorithm

The decision tree algorithm is a flowchart-like tree structure [90]. In decision algorithms, an internal node represents a feature (or attribute), the branch represents a decision rule, and each leaf node represents an outcome. The topmost node in a decision tree is known as the root node. The decision tree algorithms learn to partition data based on the attribute value. It recursively partitions the decision tree. This flowchart-like structure helps in the decision-making process like that of human-level thinking. Unlike other neural networks, decision tree algorithms share internal decision-making logic. The main advantage of decision tree algorithms is that they are faster than other neural networks. The complexity of decision trees lies in the number of records and attributes in the given dataset. The decision tree algorithms do not depend upon probability distribution assumptions. Hence, they can handle high-dimensional data with reasonable accuracy.

2.4.13 Linear Discriminant Analysis (LDA)

LDA is a generalization of Fisher's linear discriminant [91], used in statistics, pattern recognition, and machine learning to find a linear combination of features that characterizes the classes of objects. LDA classifier is loosely related to regression analysis and analysis of variance (ANOVA). However, ANOVA uses categorical independent variables and continuous dependent variables. On the other hand, LDA uses continuous independent variables and a categorical dependent variable. LDA is also closely related to principal component analysis (PCA) and factor analysis, as they both look for linear combinations of variables that match the data. LDA is used when groups are known as priori. One of the main applications of LDA is to assess the severity state of a patient and the prognosis of

disease outcomes. For example, the LDA classifier is commonly used to determine voice pathology's severity into mild, moderate, and severe forms.

2.5 Survey on Voice Pathology Detection Techniques

Voice disability detection algorithms presented in the literature can be classified based on voice features. In this section, the related research works are classified based on voice features, namely MFCCs, multiple features, time-domain features, pitch, spectrogram, and formants.

2.5.1 The MFCC Techniques

MFCCs are the most common features used in pathological voice detection. It is widely accepted that MFCCs can be used to characterize the human voice generation system fully. Hence, it is considered an effective tool for voice disability detection.

In [92], the authors develop a deep learning-based approach for the detection of pathological voices. Their work collected normal and pathological samples of eight common clinical voice disorders from a tertiary teaching hospital. The distinct pathological voices with vocal fold nodules, polyps, cysts, neoplasm, vocal fold paralysis, atrophy, sulcus, and spasmodic dysphonia are considered in the investigation. The MFCCs are extracted from the voice samples containing sustained vowel sound for three seconds and then used in three machine learning algorithms: DNN, SVM, and GMM, using five-fold cross-evaluation. To evaluate the performances of these classifiers, the authors use the voice disorder database of MEEI. The results show that the highest accuracy achieved by the DNN classifier is 94.26% and 90.52% for male and female subjects, respectively. While validating with the MEEI database, the highest accuracy of 99.32% is achieved by the DNN classifier. Based on the results, the authors conclude that having several layers of neurons and optimized weights helps DNN to outperform other algorithms.

Wavelet sub-band-based hybrid classifiers are used in [93]. Hybrid classifiers, namely GMM-UBM and Gaussian mixture model support vector machine (GMM-SVM), are used in that work. The voice samples are divided into three sub-bands using DWT. The MFCCs are computed from each sub-band. Later, the authors model the MFCCs using GMM-UBM and score them by SVM, as shown in Fig. 2.12. The results show that the accuracy of hybrid GMM-UBM for wavelet sub-band MFCCs is 96.96%, which is

significant compared to that of conventional MFCCs with GMM-UBM (i.e., 85.18%). The novelty of the proposed classifier is that it is independent of any phonemes '/a/', '/i/', and '/u/'. The proposed method considers the database of 142 normal voice samples and 147 pathological voice samples from 30-70 years. For each person, the vowels '/a/','/i/', and '/u/', each with a 1.5-second duration, are recorded at a 44.1 kHz sampling frequency. The proposed method decomposes the signal into several sub-bands using discrete wavelet transform, and then MFCCs are calculated for each sub-band. The GMM scores are extracted from each sub-band MFCC using GMM-UBM and applied as input to SVM for final classification. In the investigation, the authors use different types of wavelets. The accuracies of different wavelet types are listed in Table 2.1. It is shown that the B2 wavelet family provides the best accuracy (i.e., 92.19%). Finally, the performance matrix for GMM-UBM and GMM-SVM are recorded in Table 2.2. Based on the data, it can be concluded that GMM-SVM provides the best accuracy (i.e., 96.61%) compared to the conventional GMM-UBM.

Another MFCCs based pathological voice detection algorithm is presented in [94]. The authors focus on the classifier's capacity to improve the accuracy of voice pathology detection. They divide the classifiers into two categories, namely (a) generative (GMM and HMM) and (b) discriminative (SVM and ANN). The main advantage of generative classifiers is their capacity to handle data in separate classes. Hence, the hybrid combination of these two types is essential. The authors analyze the normal and pathological voice samples from the SVD at the Institute of Phonetics, University of Saarland, Germany. They investigate normal and pathological samples for (a) vowels with different intonations, (b) sentences, and (c) EGG sampled at 50 kHz with 16-bit resolution. The pathological voice sample considered in that work is neurological. Since this disease is more frequently seen among females, the authors only choose a female voice database. The significant findings of the work are summarized in Table 2.3 and Table 2.4. The work focuses on finding a better choice of distance metric in the radial basis function (RBF) kernel. In their work, the authors have introduced two new distance metrics, namely modified Kullback Leibler (KL) distance and modified Bhattacharyya distance (BH). They have improved 2 % and 7 % in sensitivity compared to classical KL (KL-MCS) and BH, respectively.

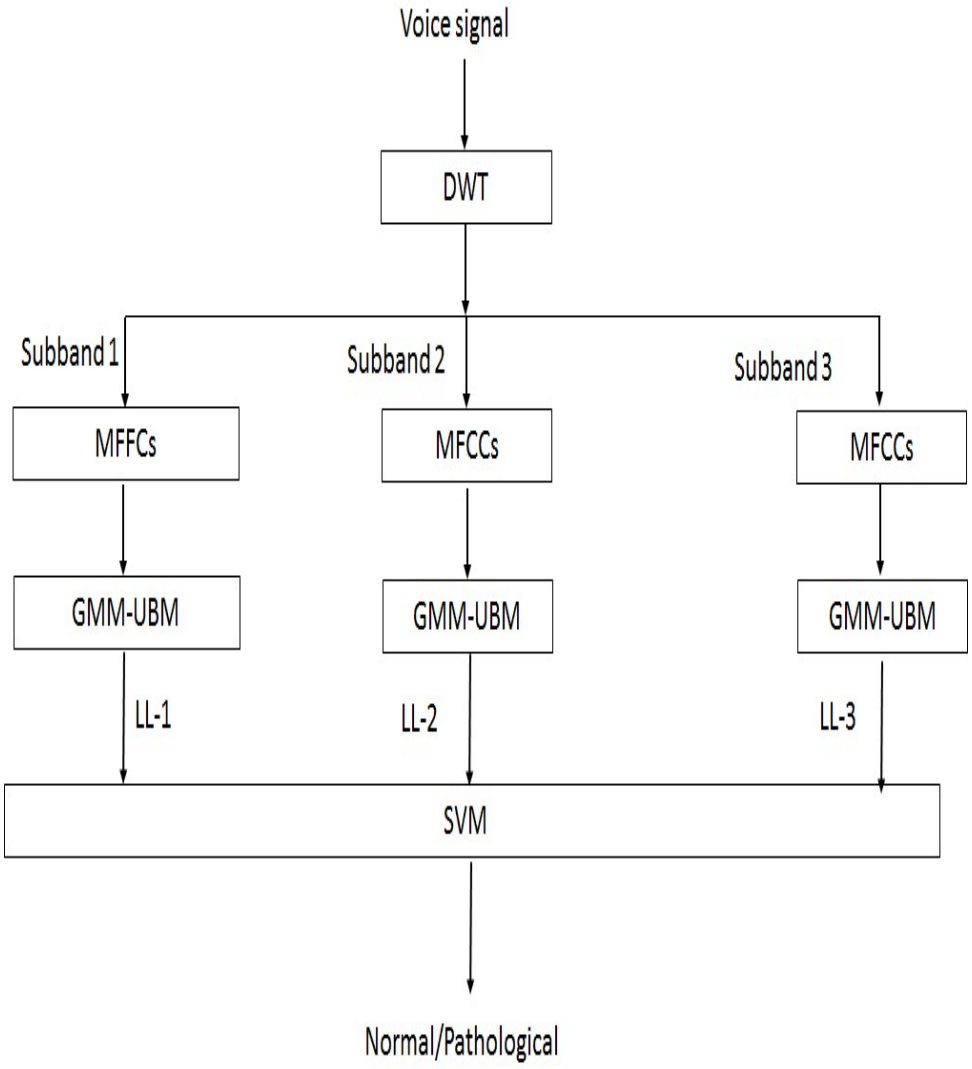


Figure 2.12 The proposed method for phoneme-independent pathological voice detection.

Table 2.1 The effects of wavelet families

Wavelets	Accuracy (%)
Haar	88.45
DB1	86.68
DB2	92.19
DB3	91.82
Symlet	91.49

Table 2.2 The performances of GMM-UBM and hybrid method

Classifier	Sub bands	Sensitivity (%)	Specificity (%)	Accuracy (%)
GMM-UBM	Full band (conventional MFCCs)	81.08	89.58	85.18
GMM-UBM	A2	82.99	93.05	88.95
GMM-UBM	D1	88.05	86.95	87.62
GMM-UBM	D2	86.05	88.34	86.15
Hybrid (GMM+SVM)	A2+D1+D2	97.19	96.00	96.61

Table 2.3 GMM-SVM results using classical and modified KL

Performance	Distances	
	KL-MCS	Modified KL
Sensitivity	92%	94%
Specificity	96%	99%
Accuracy	94%	96.5%

Table 2.4 GMM-SVM results using classical and modified BH

Performance	Distances	
	BH	Modified BH
Sensitivity	86%	93%
Specificity	96%	98%
Accuracy	92.5%	95.5%

Voice pathology due to Parkinson's disease is addressed in [95]. The proposed approach operates on cepstral features extracted from voice samples using a 30 ms Hamming window. For each signal frame, 12 MFCCs together with log-energy are calculated. The authors argue that biomedical acoustic distortions of voice signals occur during the acquisition and transmission process, and those distortions affect acoustic features extracted from pathological voice. Hence, the information about these distortions can be used to compensate for the effect. The authors propose an algorithm for detecting

four significant types of acoustic distortions in their work. The authors use GMM and LDA to detect noises. They also use two more classifiers, namely SVM and probabilistic LDA, to determine specific types of distortion in voice. The authors use clean and acoustically distorted pathological voices, achieving an 88% overall classification accuracy.

A computerized classification model is presented in [96] to diagnose vocal pathology. The authors use state-of-the-art machine learning algorithms and various classifiers in their work. The authors transfer the acoustic waveform of voice record into Mel-spectrogram and then extract features for dense net recurrent neural network (DNRNN) and feature-based classifiers. The results show that the DNRNN algorithm achieves an accuracy of 71%, the recurrent neural network (RNN) achieves an accuracy of 30%, and the random forecast approach achieves an accuracy of 68%. Based on the results, the authors conclude that frequency-domain voice features are more appropriate for detecting voice pathology than their time-domain counterparts.

In [97], the authors claim that most vocal fold pathologies cause changes in the voice signal. Therefore, voice signals can be a valuable tool for diagnosing them. The paper presents a vocal fold pathology detection technique with voice signal processing. The authors first extract MFCC voice features, then classify the feature vector using GMM. The authors also present the design and implementation of their system in that work. They show that their proposed method is less computationally complex compared to other related algorithms. The experiment was conducted on 30 speakers, and the speech duration was 60 seconds. The signal processing steps performed in that work are shown in Fig. 2.13. The preprocessing step reduces the effect of noise, removes dc offset, and performs pre-emphasis. The framing and windowing step samples the voice using a 32 ms Hamming window. The feature extraction uses a filter bank of size 12 in the frequency range of 0-8 kHz. Eleven coefficients are taken from MFCCs. In the training step, the GMM algorithm of different orders is used. In the test phase, the decision is made regarding normal and pathological samples. The performance matrix shows that GMM provides the best accuracy.

In [98], the authors argue that brain tumors, lesions, neural degeneration, and brain injury affect the speech-producing center in the human brain. Hence, the voice contains hidden information about the disorders in the nervous system. The authors use a speech

processing algorithm to detect pathological conditions of the brain. The work investigates the use of MFCC and SVM for diagnosis. The voice signal of 1.5-second duration is segmented by Hamming window of 20 ms with overlays of 10 ms. Thus, 149 frames are generated, and 13 MFCCs are computed for each frame. The authors test and train the SVM classifier using normal and pathological subjects with multiple voice disorders. They demonstrate that the accuracy level is significantly high with SVM.

A method for identifying and classifying pathological voices using ANN is presented in [99]. Several other classifier algorithms, namely multilayer perceptron neural networks (MLPNN), GRNN, and PNN, are used for classifying pathological voices. The MFCC features, extracted from audio recordings, are used for this purpose. Results show that the performance of PNN and GRNN is similar. It is also found that MLPNN performs better than PNN and GRNN in classifying pathological voices using MFCC features.

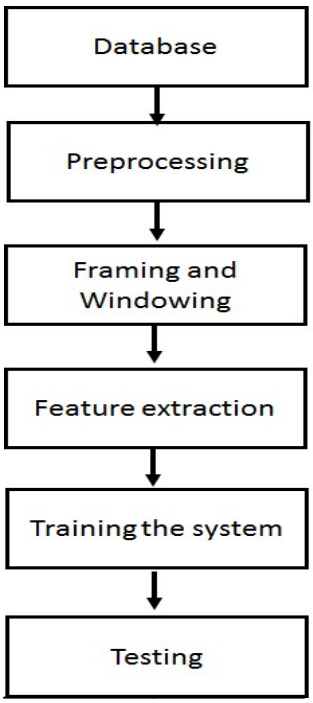


Figure 2.13 The signal processing steps used in [97].

MFCC-based voice disability detection algorithms are summarized in Table 2.5. Based on the data listed in Table 2.5, we can conclude the followings. Most of the works use ‘vowel’ samples. SVM is the most popular algorithm used in MFCC-based voice

disability detection algorithms, and the MLPNN classifier achieves the best accuracy (i.e., an accuracy of 100%).

2.5.2 Multiple Features

The primary motivation for using multiple voice features is to improve detection accuracy. The researchers show that a single feature may not detect voice pathology with high accuracy. Hence, multiple voice features can improve accuracy.

An automatic speech recognition (ASR) system called the hidden Markov model tool kit (HTK) is used in [100] for identifying pathological voices. By using HTK, the highest accuracy achieved is 94.44% for normal voices and 88.63% for pathological voices. This algorithm uses voice features, including MFCCs, PLP, RASTA-PLP, and LPC. The authors consider voice samples of 297 speakers; 121 are normal, and the remaining 176 have five types of vocal fold disorders. The results show that the best accuracy is 94.44% for MFCC with normal samples, and LPC shows the least performance of 77.25%. The other parameters show accuracies of 94.44% and 89.62% for PLP and RASTA-PLP, respectively. For pathological voice, PLP provides the best accuracy (i.e., the accuracy of 88.63%). Others are respectively MFCC (accuracy of 87.65%), RASTA-PLP (accuracy of 87.14%), and LPC (accuracy of 76.16%). The main shortcoming of this work is that the authors use manual segmentation. They use a 5-fold cross-validation in the algorithm. Arabic digits (‘0’ to ‘10’) and the Arabic words ‘ganal’, ‘gazal’, and ‘zarf’ are used for classification. The authors also present an automatic segmentation technique using fuzzy logic in the same work.

In [101], the authors present a computer-based algorithm for classifying pathological voices from normal voices. In that work, 50 voice samples are investigated (20 normal samples and 30 pathological samples). The features used include energy means, ZCR max, ZCR mean, LPC, and MFCCs for different voice segment durations of 200, 300, 400, and 500 ms. The threshold value for each feature is calculated based on the values that are best to distinguish between normal and pathological voices. The work is focused on detecting laryngeal voice disorder. The results of the work are summarized in Table 2.6. It can be inferred from the table that the highest accuracy is achieved with ZCR features, and the lowest accuracy is achieved with MFCC features.

The LPC-based cepstral analysis is used to discriminate pathological voices in [102]. The main objective is to detect vocal fold edema. The investigated voice features include cepstral (CEP), delta cepstral (DCE), weighted cepstral (WCEP), and weighted delta cepstral (WDCEP) coefficients. The vector quantization technique is used to classify normal and pathological voices. The authors consider 44 pathological voices (33 women and 11 men), most of them (i.e., 32) with bilateral edema. The normal samples considered in that work are 53 patients (21 male and 32 female). All normal voice samples are down-sampled to 25 kHz. The database contains more than 1400 voice samples with sustained vowel '/a/' from around 700 subjects. The results of the work are summarized in Table 2.7. The authors also present the ROC curve for all coefficients. The results show that DCE provides the best efficiency in pathological voice detection; however, CEP provides the correct acceptance rate.

Table 2.5 Summary of MFCC-based techniques

Research works	Samples	Phonemes	Pathological Condition	Classifier	Summary of findings
Shih-Hau Fang [92]	Normal: 60 Pathological: 402	Vowels	Structural lesions, neoplasm, neurological disorder	SVM, GMM, DNN	- SVM outperforms GMM. - DNN provides the highest accuracy
Vikram and Umarani [98]	Normal: 142 Pathological: 147	Vowels '/a/', '/i/', '/u/'	General	SVM, GMM-UBM, GMM+SVM	- Voice signal is decomposed into sub-band by wavelet transform - The MFCCs are extracted from the sub-band signals - GMM+SVM outperforms SVM and GMM-UBM, and the highest accuracy is 96.61% - Wavelet dB2 provides the best performance.
Fethi. Mohammed, and	Samples: 200	Vowels '/a/', '/i/', and '/u/',	Spasmodic Dysphonia	GMM-HMM , SVM-ANN, GMM-SVM	- GMM-SVM outperforms the other two algorithms, provided the distance metric is suitably chosen.

Hocine [94]		"Good morning, how are you?"			
Amir Hussain et. Al [95]	Samples: 3750	Vowel '/a/'	Parkinson's disease	GMM, LDA, SVM	- The method presented to detect specific noises, for example, background noise, reverberation, - clipping, and coding. - Performances of the SVM classifier can be - improved by 11% if noise information is used.
Tae Joon [96]	Not mentioned	General voice samples	Neoplasm, phono-trauma, vocal palsy	DNRNN RNN	- The highest accuracy is achieved with DNRNN
Paravena et. al [97]	Samples: 320	General voice samples	Vocal fold pathology, Coughed speech, Fan noise	GMM of order 8,16,32	- GMM of order 16 produces the highest - accuracy of 98% - GMM of order 8 produces the lowest - accuracy of 83%
Vikram [93]	Normal: 56 Pathological: 67	Vowel '/ah/'	Parkinson's disease, Vocal cord paralysis, cerebral demyelination	SVM	- Highest accuracy of 93%
V. Srinivasan [99]	Samples: 20	General voice sample	General	Multilayer Perceptron Neural Networks, Generalized regression Neural Networks	- MLPNN achieves an accuracy of 100%

A supervised algorithm is used in [103] to classify pathological voices from normal voices. The voice features considered are MFCCs and the energy variation of Jitter and Shimmer. The authors classify the data using GMM. The procedure used is illustrated in Fig. 2.14. The results presented in [103] are summarized in Table 2.8 and Table 2.9. The main focus of the work is to detect spasmodic dysphonia only. The results show that the

best accuracy achieved is with 39 coefficients, including Jitter and Shimmer, as shown in Table 2.9. The author also claims that pathology detection is more efficient with the second derivative of MFCCs.

Pathological voice detection using HMM, GMM, and SVM is addressed in [104]. The authors compare their results with previously published work based on ANN. Six characteristic parameters, namely Jitter, Shimmer, NHR, soft phonation index (SPI), amplitude perturbation quotient (APQ), and relative average perturbation (RAP) of normal and pathological voice samples, are investigated in the study. The pattern recognition algorithm is used to categorize normal voices from pathological voices. The authors discover that GMM based method can provide a superior classification rate compared to other classification methods. In the study, the authors consider cases with vocal fold diseases, namely cysts, edema, laryngitis, nodule, palsy, polyp, and glottis cancer.

Table 2.6 Accuracy, sensitivity, and specificity for single features and combined features

Feature	Condition	Sensitivity	Specificity	Accuracy
Energy mean	> 0.07	85%	70%	76%
ZCR Max	< 0.23	80%	90%	86%
ZCR mean	[0.09:0.13]	100%	67%	80%
LPC	[110:130] or [167:220]	75%	87%	82%
MFCC	[130:150]	60%	57%	58%
ZCR mean and ZCR max		80%	97%	90%
ZCR max and LPC		65%	93%	82%
ZCR max and MFCC		50%	97%	78%
ZCR means and energy mean		85%	87%	86%
Energy means and MFCC		55%	90%	76%

Detection of dysphonia is addressed in [105]. Dysphonia is a disorder that occurs when the voice quality, pitch, and loudness are altered. About 10% of the population suffers from dysphonia. It is caused mainly by unhealthy social habits and voice abuse. The authors use a mobile device to detect voice pathology. In the study, several voice features, namely MFCC, noise features, temporal derivatives, Jitter, Shimmer, wavelet transform, NHR, SPI, APQ, RAP, spectral features, perturbation, and energy parameters, are used. Several machine learning algorithms are used in that work, including SVM, decision tree,

Bayesian classification, logistic model tree, and instance-based learning. The results are compared in terms of accuracy, sensitivity, specificity, and receiver operating characteristic (ROC). The results show that the SVM and decision tree algorithm achieves the best accuracy.

The work, presented in [106], evaluates the accuracy of different characterization methods for the automatic detection of multiple speech disorders. The pathologies considered in the paper include dysphonia due to Parkinson's disease, laryngeal pathologies, and hypernasality in children with cleft lip and palate. The authors use four methods: noise content measure, spectral-cepstral modeling, non-linear features, and stability in fundamental frequency. They conclude that the stability measure is suitable for Parkinson's disease and laryngeal pathologies. The spectral cepstral features are eligible for the detection of hyper-nasal voice. Noise measures are ideal for dysphonic voices. The authors also conclude that a particular feature is not convenient equally to model all voice pathologies. Hence, it is vital to study the physiology of each impairment to choose the most appropriate set of features.

Table 2.7 Performance comparison of different cepstral methods

Methods	Correct Rejection (%)	False Acceptance (%)	Correct Acceptance (%)	False Rejection (%)	Specificity (%)	Sensitivity (%)	Efficiency (%)
CEP	89	11	91	9	89	91	90
WCEP	94	6	86	14	94	86	90
DCE	98	2	86	14	98	86	92
WDCEP	91	9	82	18	91	82	87

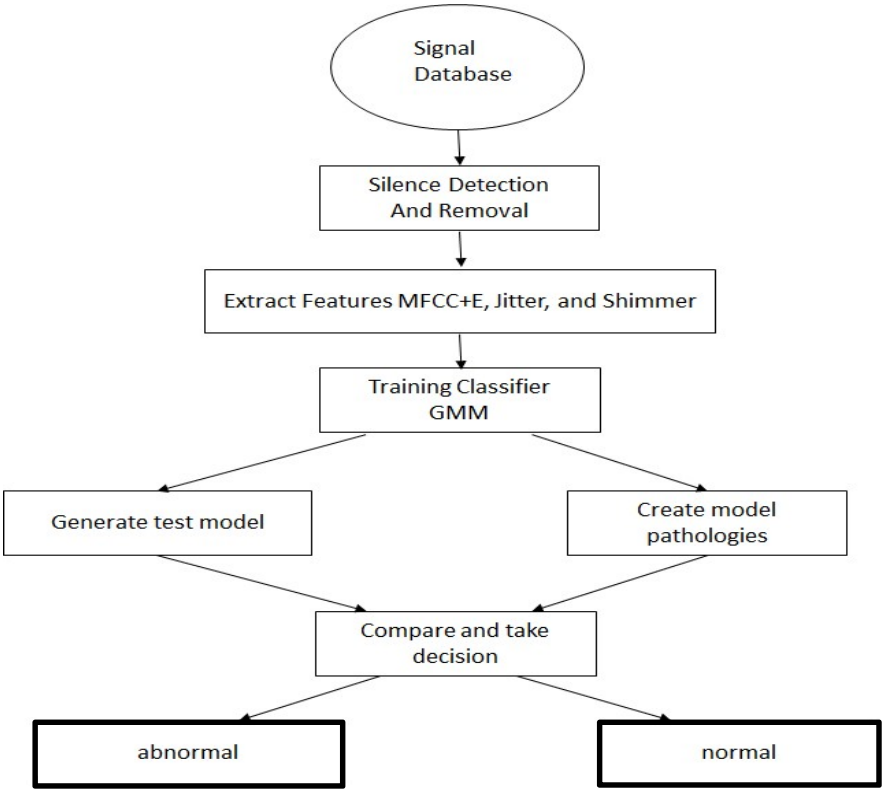


Figure 2.14 Pathological voice detection by GMM [103].

Table 2.8 The confusions matrix with MFCC and energy coefficients

System's Decision	Actual diagnosis (MFCCs and Energy)	
	Pathological (%)	Normal (%)
Pathological	79.92	18.10
Normal	20.08	81.90

Table 2.9 The confusion matrix with MFCC, Jitter, and Shimmer coefficients

System's Decision	Actual diagnosis (MFCCs, Jitter, and Shimmer)	
	Pathological (%)	Normal (%)
Pathological	82.14	17.4
Normal	17.86	82.6

In [107], Jitter, Shimmer, periodic correlation, and GNE detect voice pathology. An additional feature, namely the speech signal's noise content, is used in that work. The authors argue that GNE is an acoustic measure that has advantages over NNE or Cepstrum-based harmonics-to-noise ratio. The authors conclude that GNE is independent of fundamental frequency (Jitter) variations and amplitude. A two-dimensional "hoarse" diagram is also presented in the paper. The "hoarse" diagram can be used to determine the severity of voice disability. In the hoarse diagram, Jitter, Shimmer, and periodic correlation contribute in equal parts to the *x*-coordinates, while a linear function of GNE defines the *y*-coordinate. The authors consider that a hoarse diagram is a suitable tool for differentiating various phonation mechanisms and specific vocal pathologies and monitoring the voices' progress during voice rehabilitation.

Detection of vocal fold pathology with the aid of speech signals recorded from the patients is presented in [108]. The authors separate pathological voice from normal voice by using voice feature analysis. Their method consists of two steps. In the first step, voice features, including MFCC, LFCC, and ZCR, are extracted from the voice samples. In the second step, the classification is done by using ANN. The main advantage of the proposed method is that it has less computation and it supports real-time system development.

The work presented in [109] aims to compare and evaluate dynamic feature sets suitable for classifying pathological voices using HMM. Various features, including MFCC, HNR, GNE, NNE, and energy envelopes, have been used in that work. The feature extraction is carried out employing PCA, and the classification is done using discrete and continuous HMM. The results show that there is a direct relationship between principal direction and classification performance. The authors claim that dynamic feature analysis (employing PCA) reduces the dimension of the original feature space while keeping the topological complexity unchanged. The algorithm is tested with Kay Elemetrics (DB1) and UPM (DB2) databases. The results show that an accuracy of 91% can be achieved from the proposed algorithm with a 30% computational cost reduction for DB1.

The work presented in [110], explores and compares various classification models to find acoustic parameters' ability to differentiate normal voices from pathological voices. The authors use different classification algorithms, namely SVM and radial basis function neural networks (RBFNN). Acoustic parameters include signal energy, pitch, formant

frequencies, mean square residual signal, reflection coefficients, Jitter, and Shimmer. The acoustic features are combined to form a feature set. The results show that the RBFNN algorithm can achieve an accuracy of 91% compared to 83%, which can be achieved by SVM.

Stuttering voice disability is addressed in [111]. The authors claim that over 3 million American stutters when they speak, and many voice interfaces that exist with consumer technology often neglect the population with voice ailment, including TV and car systems. For example, Apple's Siri is tested against various speech disorders, including stutter and slurred speech. It is found that accuracy ranges from as low as 18.2% to only as high as 73%. The authors propose a method to improve the performance of automatic speech recognizers on speech containing stuttering. Specifically, the authors develop a classifier that can better detect stuttering in speech signals and study techniques for applying these classifiers to automatic speech recognition mode. The classifiers can effectively parse out stuttered speech before processing the same. The classification algorithm includes ANN, HMM, and SVM. The authors implement a six-layer neural network algorithm using MATLAB.

A system for remotely detecting vocal fold pathology using telephone-quality speech is implemented in [112]. The system uses a linear classifier to process the measurement of pitch perturbation, amplitude perturbation, and harmonic-to-noise ratio derived from speech samples. The results show that an accuracy of 89.1% can be achieved when the voice is recorded in a controlled environment—however, the same declines to 74.2% when telephone-quality speech is used. The authors classify voice samples into four subclasses: normal, neuromuscular, physical, and mixed (neuromuscular and physical). The significance of this study is that it combines telephony and server-side speech processing to diagnose pathological voices from a remote location.

In [113], the authors argue that pathological voice detection algorithms often fail to detect voice pathology correctly. Additionally, classification rates are still insufficient for reliable and large-scale screening. The work reviews the performance of state-of-the-art methods and their weaknesses. The authors include the features in the time and frequency domain. Different machine learning techniques evaluate the features. Based on the results, they conclude that the spectral features are the most important. On the other hand, pitch-

related features are less important. The most useful feature set is the residual from the inverse LPC filtered signal. The authors also show the effectiveness of their algorithm.

In [114], the authors investigate dysphonic voices. Sustained vowels from male and female speakers with mild to severe dysphonia are analyzed in that work. Multiple voice features are used. The authors make several essential conclusions in the paper as follows. The reliability of F_0 measurement decreases significantly with increasing dysphonia. The shimmer measures vary much more in reliability at all levels of severity than F_0 measures and the reliability is not related significantly to increasing dysphonia. The overall reliability is even worse for Jitter and HNR than for F_0 and Shimmer.

Five voice qualities have been used in [115] to detect vocal disorders. The work also investigates six acoustic measures. The authors extract all the measures from the residue signal obtained by inverse filtering the speech signal using the LPC technique. The authors conclude that pitch amplitude (PA) and HNR are the two most valuable parameters for predicting vocal quality.

Grade Roughness Breathiness Strain (GRBS) attributes of voice have been presented in [116] to assess pathological voice quality automatically. The proposed method adopts higher-order local autocorrelation (HLAC) features. The HLAC features are calculated from the excitation source signal obtained by an automatic topology-generated autoregressive higher-order HMM (AR-HMM) analysis. Additionally, the proposed method identifies the four attributes using a feed-forward neural network (FFNN) based classifier.

In [117], the authors argue that although there are many research works published to detect pathological voices; however, only a few deal with the severity of estimation of voice disabilities. The authors present an automatic classifier using acoustical measurement of sustained vowel '/a/' and a pattern recognition tool based on neural networks. The authors include four acoustic features. The severity of the voice disability is estimated depending on how these parameters are far from standard values. The authors use healthy and pathological voice samples from a German database in the analysis. The performance of the proposed algorithm is evaluated in terms of accuracy (97.9%), sensitivity (1.6%), and specificity (95%). The results show that the classification rate is 90% for the normal class and 95% for the pathological class. The authors recommend

threshold values of Jitter, Shimmer, and HNR, as listed in Table 2.10, to differentiate between pathological and normal voices.

The research presented in [118], compares the effectiveness of pitch rate, Jitter, Shimmer, and HNR as indices of voice disability in English, German, and Japanese language speakers. This study includes reciting a page instead of using only long vowel sounds. The results show that Jitter, Shimmer, and HNR are effective indices for long English vowel sounds. On the other hand, Shimmer and HNR for reading speech are considerably worse, although the effectiveness of Jitter is an index that is maintained for reading speech. The pitch rate is better in distinguishing healthy individuals from patients with illnesses affecting their voice. The reading speech results in German, Japanese, and English are similar.

In a study in [119], the authors present a robust, rapid, and accurate system for detecting normal and pathological speech automatically. The system uses fully automated measures of vocal tract characteristics and excitation information. The authors use MFCC coefficients and pitch dynamics to model the Gaussian mixture in HMM classifier. The authors compare their work with some existing best-performing work, demonstrating that their method outperforms other classifiers by 8%. The authors use two methods: multi-dimensional voice program (MDVP) and HMM. The results are summarized in Table 2.11 and Table 2.12. These two tables show that GMM provides the highest accuracy using MDVP. However, the accuracy is 99.44% when MFCC and pitch are combined.

Table 2.10 The recommended ranges of the parameters for voice disability detection

Software		PRAAT	Teixeira
Jitter (ddp%)	Female	≤ 1.04%	≤ 0.66
	Male	≤1.04%	≤0.44
Shimmer (dda)	Female	≤ 3.810	≤2.43
	Male	≤3.810	≤2.01
HNR (dB)	Female	≤ 20dB	15.3 dB
	Male	≤20 dB	17.3 dB

The summary of the mixed features-based voice disability detection algorithms is presented in Table 2.13. Based on the table, we can conclude the following. Although more

than one voice feature has been used, MFCC is one of the most common features. Mainly vowels are used for generating voice samples. Among the classifiers, SVM and ANN are commonly used in multiple feature-based algorithms.

Table 2.11 The classifications using MDVP

Method	Training (%)	Test (%)
LDC	95.64	95.93
NMC	67.15	65.24
GMM	97.97	97.67

Table 2.12 The classifications using HMM with MFCC and MFCC + pitch

Features	Training (%)	Test (%)
MFCC	98.59	97.75
MFCC+ Pitch	99.44	98.30

2.5.3 Time Domain Features

In voice disability detection algorithms, voice features, other than the time domain, have been mainly used. However, some recent works show that time-domain parameters can also be used effectively in voice disability detection. Some of these works are now presented.

Table 2.13 The summary of mixed features-based classifications

Research	Samples	Phonemes	Pathological Condition	Features	Tools	Summary of findings
Manu Chopra [111]	28 samples	3 minutes speech samples	Stuttering, Slurred speech	MFCC, Spectral measures	ANN, HMM, and SVM	-Stuttered voice can be improved by 87% for male and 75% for female.
R.J. Moran et. al [112]	Normal: 58 Pathological: 573	1-3 second speech samples	Neuromuscular, Physical, and both	Fundamental frequency, Shimmer, Jitter, Perturbation in amplitude, SNR, and HNRs	LDA	-Obtained accuracies :87% neuromuscular, :78% abnormality, and :61% mixed pathology
Zvi Kons [113]	Samples: 320 males and 339 females	Vowel '/a/' 2-5 second	Nodule, Polyp, Cyst, Cancer, Polypoid, Hyperplasia, Keratosis, and Papilloma	Pitch, Degree of voicing, Spectral envelope, Harmonic frequency, Jitter, LPC, LPC residual signal, and Glottal sound	SVM	- Severe cases are easier to diagnose and weak pathology is hard to diagnose.

Algarbi in [100]	Normal: 121 Pathology: 176	Arabic words '/gamal/' '/gazal/' '/zarf/'	Irregular vibration and Incomplete closure of vocal fold	MFCC, PLP, RASTA-PLP, and LPC	HTK	-Accuracy of 94.44% for normal voice and 88.63% for pathological Voices
Stevan [114]	Male: 29 Female: 21	Vowel '/a/'	Dysphonia	Fundamental frequency, Jitter, Shimmer, Harmonics, signal to noise ratio, HNR	CSpeech, Computerized speech laboratory (CSL), and Soundscope	-Wide variation of reliability with increased pathology using shimmer
L. Eskenazi [115]	Male: 25 Female: 25	Vowel '/a/'	Hoarseness, Breathiness, Roughness, and Vocal fry	HNR, PA, and JIT	Prediction Sum of Squares (PRESS)	- The most useful parameters for voice pathology detection are: (a) Overall Quality: PA and HNR. (b) Breathy voice: SIR and HNR (c) Vocal fry: PA and HNR (d) Hoarse Voice: PA and HNR
Akira [116]	Pathological: 60	Japanese Vowel	Roughness, Breathiness, Asthma, and Strain	HLAC	FFNN, AR-HMM	- An accuracy of 87.75% is achieved to detect voice pathology.
Brahim et al [117]	Normal: 25 Pathological: 25	Vowel '/a/' 3-5 second	General	Pitch, Jitter, Shimmer, and HNR	ANN	- Acoustical measurement is helpful to detect the severity of pathology.
Shuji [118]	Normal: 53 Pathological: 602	Vowel '/ah/', Rainbow passage (German, Japanese, and English)	Hyper function, Paralysis, Anterior-posterior squeezing, Gastric reflux, Vocal fold edema, and Ventricular compression	Pitch, Jitter, Shimmer, and HNR	PRAAT	- Voice pathology detection depends on the language- more efficient for English, but less efficient for German and Japanese.
Alireza [119]	700 samples MEEI	Vowel '/a/' Rainbow passage	Organic, Neurologic, Traumatic, and Psychogenic	MFCC, Pitch	GMM, MDVP	- An accuracy of 99.44% is achieved.

The focus of the work presented in [120] is to detect voice disability among children. In that work, the authors use the envelope of voice signals to detect pathological cases of speech-disabled children. The speech samples of children aged 5-8 years are used in the study. The speech signals are first digitized, and then speech envelopes are detected. The envelopes are then used for ratio mean analysis to estimate speech disability. The authors also classify voice disability into three levels.

It is claimed in [109] that the short-term parameters combined with dynamic classifiers such as HMMs are suitable for the pathological voice detection system. The authors argue that most approaches rely on complex procedures or add new parameters that

increase the processing time and do not favor the system's performance. They present an approach that improves the standard scheme of HMM-based classifiers to detect voice pathologies. The authors use HMMs to derive discriminative voice features defined by specific components. The authors achieve high accuracy using a simpler procedure to generate an optimal decision boundary. The results show that the proposed system significantly outperforms other classification systems.

In [121], the authors use a TEO feature to detect normal and pathological voices automatically. The authors get the idea of TEO from a work that used the LP residual phase for speaker recognition. The authors use a second-order polynomial classifier on a subject. They also use two different methods: the TEO phase and score level fusions. The comparison of the two methods is listed in Table 2.14 in terms of classification accuracy (ACC) and equal error rate (EER).

RASTA-PLP is used in [122] to identify four different types of vocal fold disorders. This study investigates dysphonic patients consisting of 40 males and 20 females. The diseases are classified by using a multi-class SVM. RASTA-PLP voice features are first extracted from the voice samples. Then the voice features are compressed using a vector quantization, which is implemented using the K-mean algorithm. The results show that a 100% classification rate can be achieved by choosing a suitable word for each disease.

Table 2.15 presents the summary of time-domain features-based voice disability detection algorithms. The vowels and other native words have been used for generating voice samples. Based on the data presented in the table, we can conclude the followings. The highest accuracy obtained is 100% for a specific pathology.

Table 2.14 The comparison between the TEO phase and score level fusions

Feature Dimensions	TEO Phase		Score-Level Fusion	
	ACC	EER	ACC	EER
6	80.65	19.34	97.50	2.49
12	79.87	20.13	97.32	2.68
30	82.66	17.23	97.28	2.71

Table 2.15 The summary of time-domain features

Research works	Samples	Phonemes	Pathological Condition	Features	Tools	Summary of findings
Anandthirtha [120]	Normal: 60 Pathological: 13 (5-8 years)	Kannada Words 'Namma' 'Nanna' 'Ide' Shale' 'Jep' 'Hesa' 'Naga' 'Saha' 'Noora' 'Jayn'	General	Envelope detector	Threshold values	- Classify the voice disability into mild, moderate, and severe.
M. Sarria [109]	Pathology: 65 Normal: 13	Spanish vowel	Dysphonia, Hypernasalit Dysarthria	Nonlinear parameter, entropy	HMM tools	- Accuracy of 99% is obtained.
Hemant [121]	Pathological: 173 Normal: 53	Vowel '/ah/'	Paralysis	Teager Energy Operator (TEO)	HMM tools	-The maximum accuracy for selection fusion is 97.28%, and TEO Phase is 82.6%
Mansour [122]	65 samples	Vowel '/a/'and '/i/'	Cyst, gastroesophageal reflux disease (GERD), Polyp, and Sulcus	PLP, RASTA-PLP	SVM	-100% accuracy in classifying GERD and polyp. Maximum accuracy of 75% and 83% for cyst and sulcus respectively.

2.5.4 The Pitch

Pitch is another essential feature used in voice disability detection. In the past, the pitch was considered an effective tool for voice recognition. Nowadays, many voice pathology detection algorithms have used pitch too.

In [123], the authors present a new pitch detection algorithm suitable for detecting dysphonia voices. The proposed method uses the frame size of half-wave rectified autocorrelation adjusted to a smaller frame after two potential pitch candidates are identified within the preliminary frame. This method is compared to PRAAT's standard autocorrelation tool, and the results show a significant improvement in detecting pitch for pathological voices. The technique is more reliable for detecting pitch in a low or high-pitched voice without adjusting the window size. The authors argue that PRAAT works better for normal voices. However, the results shown in the paper dictate that PRAAT works poorly for pathological voices. The results also show that, in some cases, PRAAT exceeds 40% of error, but their proposed algorithm never exceeds 40% of error.

A software system for pathological voice analysis using a personal computer with a sound card is presented in [124]. The software system can evaluate pitch period, degree of unvoiceness (DUV), pitch perturbation quotients (PPQ), amplitude perturbation (APQ) quotients, dissimilarities in surfaces of the pitch pulses (DPP), the ratio of aperiodic/periodic components in cepstral energy (APR), HNR, degree of hoarseness (DH), the ratio of cepstral energies (PECM), and glottal closing quotient (GCQ). The results show that the software can detect pathological voices by using the above-mentioned voice features.

Unlike other works, the unvoiced part of voice samples is investigated in [125]. The authors argue that most of the existing works depend on the voiced portion of a speech sample to detect voice pathology, and these works use pitch detectors to separate the voiced part from the unvoiced part. However, the existence of voice pathology affects the speakers' vocal fold and produces a more irregular vibration pattern. All these consequently cause degradation of voice quality within less voiced segments. Hence, selecting only clear-voiced segments for the classifier may not be appropriate. In the paper, the authors propose a new approach that enables the classification of voice pathology by analyzing the unvoiced information of the continuous speech. The signal frames are

divided into turbulent or non-turbulent instead of voiced/un-voiced parts. The results show that useful pathological information is present in turbulent or near-unvoiced segments.

The summary of the pitch-based voice disability detection algorithms is presented in Table 2.16. Based on the data, we can conclude the followings. The pitch feature is very useful for detecting voice disabilities. Although PRAAT is widely used in voice pathology detection; however, some algorithms in the time domain can provide even better accuracy for dysphonia, laryngeal, and neurological voice disorder. The vowels are mostly used in the analysis.

2.5.5 The Spectrogram Features

The spectrogram is computed based on frequency domain information. In many pathological voice detection algorithms, a spectrogram of the voice signals solely has been used.

Pathological voice disorder due to vocal cord paralysis or Reinke's edema is investigated in [126]. In the paper, the authors claim that the deep learning method is widely used in speech recognition; however, it can also be applied in pathological voice detection. The authors use CNN in the work, instead. The spectrograms of pathological and normal speech are computed and used as the input to the convolutional deep belief network (CDBN) to train CNN. Then, CNN is trained using a supervised back propagation learning algorithm to fine-tune the weights. The signal processing steps used in the work are shown in Fig. 2.15.

In [127], the authors argue that the most used acoustic measures for the diagnosis of voice disability are Jitter, Shimmer, and harmonics-to-noise ratio. However, these measurements are not independent and, therefore, may give ambiguous information. For example, the addition of random noise increases Jitter measurement, and the introduction of Jitter causes a reduced harmonic-to-noise ratio. The authors suggest that to increase accuracy in detecting voice pathology by analyzing spectrogram, it is required to remove the effects of Jitter and Shimmer on the speech spectrum. The authors test their algorithm by initially moving them on specially designed synthesis data files.

The spectrogram-based voice disability detection algorithms are summarized in Table 2.17. We can conclude the following based on the data. Vowels are primarily used

as voice samples. The deep learning algorithm helps detect voice disability. Jitter and Shimmer adversely affect the voice disability decision.

2.5.6 The Formants

Like spectrogram, formants are also a frequency domain feature. It has been widely used in voice recognition algorithms. However, some voice pathology detection algorithms have used formants as the primary tool.

Table 2.16 The summary of pitch-based voice disability detection algorithms

Research works	Samples	Phonemes	Pathological Condition	Features	Tools	Summary of findings
Mohammad Redzuan et al [123]	Normal: 49 Pathological: 87	Vowel '/a/'	Dysphonia	Pitch	Pitch Detection Algorithm (PDA)	- The proposed algorithm performs better than PRAAT in detecting voice pathology.
Boynov [124]	Normal: 100 Pathological: 300	Vowel '/a/'	Laryngeal, Neurological	Pitch	ANOVA	- Very significant changes in DH, DPP, DUV, APR, and PECM
F. Perdigae [125]	Normal: 53 Pathological: 660	Vowel '/a/' and Rainbow passage	General	Pitch	Multilayer Perceptron	- The highly turbulent speech contain useful pathological information.

The first two formants of vowels are used in [128] for voice disorder classification. Five voice disabilities are addressed in the work. The four voice features are used by two types of classifiers, namely vector quantization and neural network. The results show that neural network performs better than vector quantization in terms of accuracy.

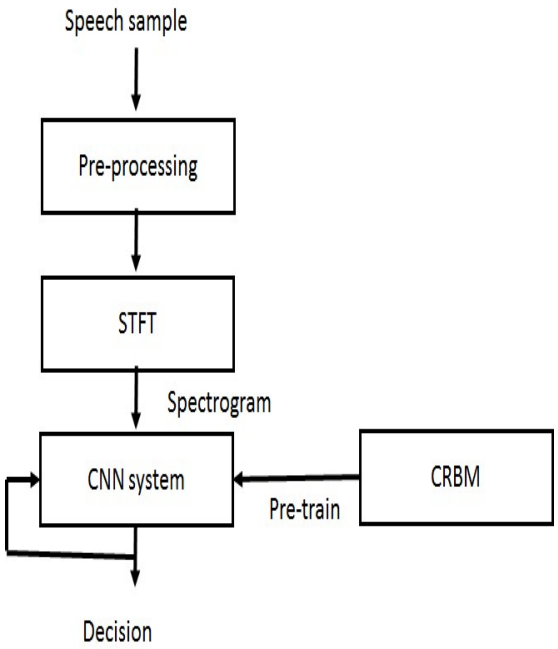


Figure 2.15 The signal processing steps used in [126].

Table 2.17 The summary of spectrogram based voice disability detection algorithms

Research works	Samples	Phonemes	Pathological Condition	Features	Tools	Summary of findings
H. Wu [126]	Pathological : 73	Vowels '/a/', '/i/', '/u/', 'Good morning, How are you?'	Reinke's edema, Laryngitis, Leukoplakia, Recurrent laryngeal, Nerve paralysis, vocal fold carcinoma, and Vocal fold paralysis	Spectrogram	CNN, CDBN	- Deep-learning algorithms can be trained with a small amount of data.
Peter Murphy [127]	Normal: 12 Pathological : 13	General	Breathy voice, Vocal fry, Modal voice, Murmur, Creaky voice, and Stiff voice	Spectrogram	MATLAB	- Effects of Shimmer and Jitter need to be removed before feature extraction for best accuracy.

Four fundamental frequencies (F_0) and two F_0- independent measures are used to quantify pathological voice [129]. Two of F_0-dependent measures are computed in the time domain, and two others are computed using spectral information from a vowel. The F_0-independent measure is based on LP modeling of vowel samples. The results show that the measures on the LP model are much superior to other measures. The authors conclude that LP modeling approach to quantify vocal noise is attractive for several reasons as follows. The LP model is known to be a good model for normal voice speech. As a result, it is applied in many speech processing applications, including speech coding, speech recognition, and speech synthesis. The LP modeling is F_0-independent. This eliminates the need for computationally intensive high precision F_0 extraction algorithm. The LP model is sensitive to the presence of noise. Thus, the presence of vocal noise is reflected in the LP model output, which can be used as an indicator of voice pathology.

The summary of the formant-based voice disability detection algorithm is presented in Table 2.18. Based on the data, we can conclude the followings. The best accuracy achieved is only 70.72%, which is less than other voice features-based algorithms. The formants help detect multiple voice pathologies, including vocal noise, Cyst, Polyp, Gerd, voice Paralysis, and Sulcus.

Table 2.18 The summary of formants-based voice disability detection algorithms

Research works	Samples	Phonemes	Pathological Condition	Features	Tools	Summary of findings
Vijay Persa [128]	Normal: 53 Pathological: 175	Vowels	Vocal noise	Formants	LP	- The vocal noise is related to LP model output.
Gulam Mohammad [129]	Male: 51 Female: 51	Arabic word ‘fathma’and ‘kasra’	Cyst, Polyp, Gerd, voice Paralysis, and Sulcus	Formants	Vector Quantization and ANN	- The best accuracy achieved is 70.72%

2.6 Issues and Challenges of Voice Disability Detection Algorithm

Voice disability detection is usually initiated by using a screening method after receiving concerns from patients, parents, teachers, and healthcare service providers. During the screening, any deviation from normal voice is detected by the physicians. Vocal

characteristics, including respiration, phonation, and resonance, are investigated during the screening process. If any deviation is detected, a comprehensive assessment is followed. The typical components of comprehensive assessment include case history, oral-peripheral examination, assessment of respiration, and auditory perceptual assessment. Voice quality is assessed by examining the voice features, including roughness, breathiness, strain, pitch, loudness, and overall severity. In addition, other voice features including MFCC, spectrogram, formants, wavelets, LPC, PLP, RASTA-PLP, Jitter, Shimmer, GNR, HNR, CHNR, NNR, ZCR, LFCC, and TEO are popularly used in voice pathology detection. From voice sample collection and assessment to the final classification stage, the following issues need to be primarily considered.

2.6.1 Sample Collection Environment

Voice samples must be collected and assessed in a controlled environment [130]. It is suggested that voice data must be collected in a quiet environment. The other requirements are (i) a microphone with a sensitivity of -60 dB should be used, (ii) mouth to microphone distance should be around 10 cm, (iii) sampling frequency should be 20-100 kHz, and (iv) recording should be done in a sound-treated room with the ambient noise of less than 50 dB, and (v) microphone must be aligned 45^0 with respect to mouth.

2.6.2 Voice Samples

There is no consensus about the most representable voice samples. However, most voice detection algorithms use vowel samples. The rest of the works use sentences and running speeches. The followings are recommended in [130]

- ***Sustained vowels***

  Two vowels, namely '/a/' and '/i/,' shall be used. The vowel '/a/' is considered a lax vowel. On the other hand, the vowel '/i/' is tense. It is also recommended that the patients should be asked to say the vowel '/a/' for a sustained period of 3-5 seconds. Then, the patient should be asked to say the vowel '/i/' for a similar sustained period.

- ***Sentences***

  The sentences used in voice sample collection should be carefully designed so that they can elicit various laryngeal behaviors. For example, the following six

sentences have been recommended in [130] : (a) the blur spot is on the key again, (b) how hard did he hit me? (c) we were away a year ago, (d) we eat eggs every Easter, (e) my mama makes lemon jam, and (f) Peter will keep at the pack. The first sentence contains all the vowels in the English language. The second sentence emphasizes '/h/.' The third sentence is all voiced. The fourth sentence elicits a glottal attack. The fifth sentence elicits a nasal sound, and the sixth sentence is mostly voiceless. In addition to these sentences, other works have used the "Rainbow Passage" [131] for voice disability detection. The specialists use this passage to diagnose a patient who has vocal cord paralysis or vocal cord paresis. This passage is considered suitable to assess the mobility of the vocal cords of a patient.

- ***Running speech***

  The clinicians urge the patients to answer some standard interview questions for at least 20 seconds, such as "Tell me about your voice problem", or "Tell me how your voice is functioning." The patients are also sometimes asked to tell a simple story.

2.6.3 The Data Source and Samples

The common sources of voice samples are the local clinics. One of the primary sources of the database is MEEI voice disorders database. However, the voice recording environment and voice recording techniques are not mentioned in the database. Hence, these are important aspects that need to be considered when implementing voice pathology detection algorithms.

2.6.4 Sample Size

The data sample size also varied widely in different works. It is shown in the paper that some works have used a few samples; however, other works have used a substantial sample. For example, only a few voice samples (i.e., 20) are used in [99]. On the other hand, large samples (i.e., 3750) are analyzed in [95]. Although using large samples for training and classification is recommended, there are no general recommendations about the sample size.

2.6.5 Voice Features

Voice features, namely MFCC, spectrogram, formants, wavelets, LPC, PLP, RASTA-PLP, Jitter, Shimmer, GNR, HNR, CHNR, NNR, ZCR, LFCC, and TEO have been used in the research works. It is mostly recommended that frequency domain voice features are more helpful for detecting voice disability. However, some researchers also argue that time-domain features are more helpful for detecting voice disability in [109], [120]-[122].

2.6.6 Classification Algorithms

Several classification algorithms have been used by researchers. Some of them include SVM, GMM, GMM-UBM, SVM-Universal Background Mode (SVM-UBM), HMM, ANN, DNN, CNN, PNN, DBN, GRNN, Bayesian classifier, the K-mean clustering, the decision tree, and linear discrimination. In addition, other common tools used are HMM tool and PRAAT software. Among these algorithms, SVM is the most popular classifier algorithm that has been widely used in voice disability detection.

2.6.7 Voiced or Unvoiced

Most voice disability detection algorithms use the voiced part of speech samples. It is shown that the voiced part of speech samples elicits the glottal structure. However, some works suggest that unvoiced portions of speech are useful. Because the pathological voices are noisy and hence, they should be used as samples to correlate the voice pathology.

2.6.8 Voice Pathology

Most of the works presented in this Chapter are suitable for detecting a particular voice pathology. Only a few works deal with more than one type of voice disability. It is also recommended that the algorithm development must target a specific voice disability, not all types of disability simultaneously.

2.7 Conclusion

This Chapter presents a survey work on voice disability detection techniques available in the literature. It is shown in the literature that voice disability detection is very challenging work because the voice signal is complicated to analyze. The voice signals widely vary depending on the disability type. There have been many algorithms reported in the literature. However, none of these algorithms is suitable for detecting any specific kind of

voice disability. Hence, it is essential to target a particular disability while designing the algorithm. In this Chapter, it is also shown that choosing the voice samples is also challenging. The researchers should focus on voiced as well as unvoiced components of the samples since there is also evidence of pathology detection in the unvoiced part of the speech samples. The survey also found that a single letter, word, or full sentence with a pause can be used as voice samples. While using a complete sentence, some extra consideration should be given to transitional words and pauses. Though many databases are available as voice sources, some researchers can also collect samples according to their pathology detection criterion in a controlled environment. But during sample collection, the researchers should take extra precautions, as mentioned in this Chapter. Selecting features from the samples is the next challenge for the researchers. From the literature survey, it is concluded that most of the researchers are more confident in using the features in the frequency domain though few researchers also rely on time-domain measures for specific pathology. Using acoustic features is also not uncommon. However, it is a long-time measurement that can be sensitive to the pathological status of the patient. Multiple features analysis is also a common practice, as seen in the survey. Many classification algorithms have been used by researchers. SVM is considered the most suitable tool for voice disability detection among these classification algorithms. However, the SVM algorithm is not particularly ideal for categorizing levels of voice disability. To achieve good accuracy in classification, a large data set is required to train the classifiers and test the algorithms. Some researchers also use different tools for classification, as found in the literature. Hence, the limitation arises when there is a need to detect the level of voice disability. To design an efficient voice pathology detection algorithm, researchers must focus on the selection of proper voice samples and appropriate feature collection. Above all, they should focus on the design of a level-based voice pathology detection algorithm suitable for a distinct pathology.

CHAPTER 3
VOICE PATHOLOGY DETECTION WITH ELECTROGLOTTOGRAPHIC (EGG) AND SPEECH SIGNALS

This Chapter presents a convolutional neural network (CNN) based automated noninvasive voice pathology detection system. The proposed system functions in two steps. First, it discriminates pathological voices from healthy ones, and then, it classifies the discriminated pathological voices into one of the three pathologies. Two CNNs are used for these purposes; one works as a binary classifier to identify pathological voices. The other one works as a multiclass classifier for categorizing voice pathologies. The main objective of this work is to investigate the effectiveness of electroglottographic (EGG) and speech signals in detecting and classifying pathological voices using sustained vowel ('/a/') samples. EGG signals can assess the vibratory pattern of the vocal folds during voiced sound. On the other hand, the speech signals add spectral color to the EGG signals. Hence, their contributions to pathology identification and segregation differ, as demonstrated in this Chapter. The SVD is used in this investigation. The contributions of this investigation are three-fold:

1) The proposed system works with a small number of datasets. Hence, this automated voice pathology detection system is fast and clinically viable.
2) This system can extract the features from the raw EGG and speech signals, thus preserving most of the pathological information within the datasets. The proposed system would avoid losing pathological information that is incurred in other systems due to preprocessing and filtering stages.
3) The proposed approach investigates EGG and speech signals separately to compare and correlate their contribution to pathological voice identification and classification in terms of statistical parameters and clinical reasoning.

3.1 Related Background

As mentioned in Chapter 1, vocal folds are significant components of the human voice generation system. The area between the vocal folds is called the glottis, as shown in Fig. 3.1. This area changes depending on the voicing, unvoicing, and breathing conditions. During phonation/voicing, the vocal cords are tensed and closer; consequently, the glottis looks like a slit, as shown in Fig. 3.1(a). While during breathing conditions, the glottis becomes a narrow wedge shape, as shown in Fig. 3.1(b); however, during forced respiration, it becomes a wide triangular shape, and the

vocal cords are as far apart as possible. During pronunciation, the vocal cords vibrate and produce a buzzing sound that makes up the human voice.

With these varying activities mentioned above, the airflow velocity at the glottis becomes a function of time, as illustrated in Fig. 3.2, and the airflow velocity roughly follows the time-varying area of the glottis. With the vocal fold in a closed position, the flow begins slowly, builds up to a maximum, and then quickly decreases to zero when the vocal folds abruptly close. The whole process can be divided into three phases: closed, open, and return, as shown in Fig. 3.2. The time interval during which the vocal folds are closed, and no flow occurs, is the glottal closed phase. The time interval during which there is nonzero flow, and the airflow velocity reaches the maximum, is the open phase. Finally, the return phase is the time interval during which the airflow velocity decreases from maximum to minimum. The process repeats periodically as a series of pulses that produce "modal" voiced speech.

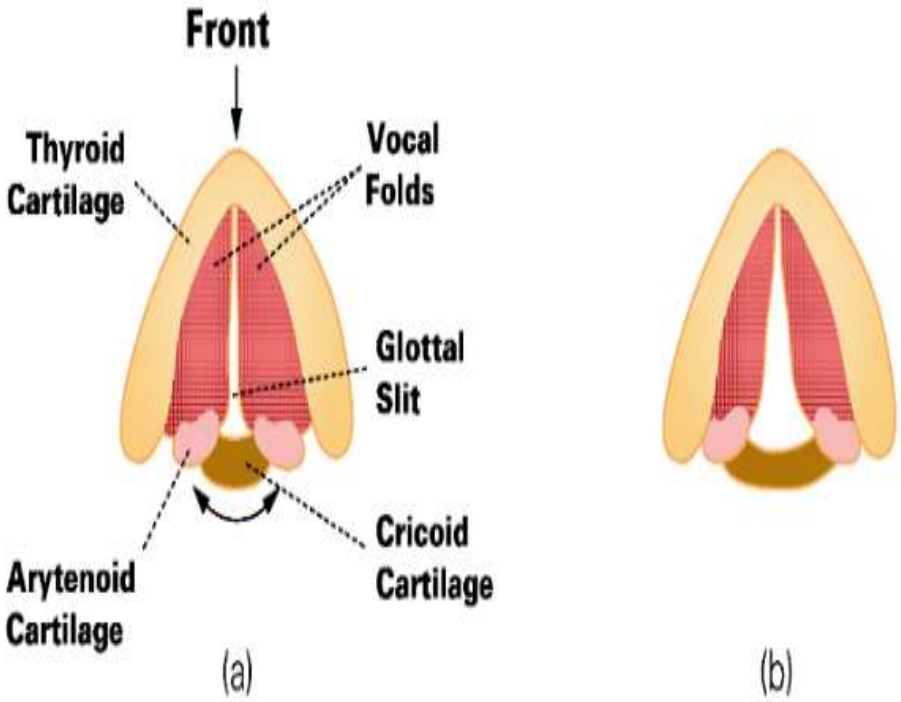


Figure 3.1 Vocal fold condition during (a) voicing, and (b) breathing [2]

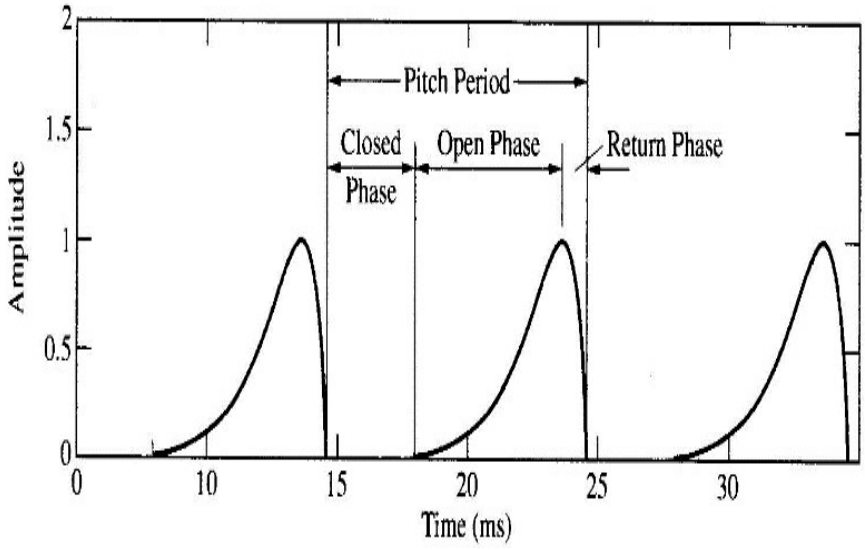


Figure 3.2 The periodic glottal airflow velocity [2].

The vocal tract output can be modeled by the convolution of the glottal airflow, $g(n)$, with the vocal tract impulse response, $h(n)$. Mathematically, the vocal tract output can be expressed as [2] :

$$y[n] = g[n] * h[n] \,, \tag{3.1}$$

where $y[n]$ and $g[n]$ are practically the speech and EGG signals, respectively.

3.2 Selection of Samples and Pathologies

In this investigation, the sustained phonation of the vowel '/a/' is used. The main reason is that a speaker can maintain a steady frequency and amplitude at a comfortable level [132], during this vowel sound generation. Moreover, it is free of articulatory and other linguistic confounds that often exist with everyday speech tasks such as sentences and running speech. Most importantly, sustained vowel sound adequately mimics the EGG waveform being free from articulation. There are several publicly available databases, some of which are as listed in Table 3.1 [133]. This study needs both the EGG and speech signals of the same patient that are available only in the SVD database [134] . Not only that, but the SVD database also meets the required sample collection criteria mentioned in [15]. These criteria are as follows: (a) the audio sample recording should be done with 16-bit resolution, and (b) the sampling frequency should not be less than 20 kHz.

Table 3.1 The publicly available voice databases [133]

Database	Samples	Language	Pathologies	Contents
MEEI [135]	Control: 53 Pathological: 657	English	Various types	Vowels, Sentences
SVD [134]	Control: 650 Pathological: 1320	Germany	71 types	Vowels, Sentences
PdA [136]	Control: 239 Pathological: 200	Spanish	Various types	Vowels
PDS [137]	Control: 8 Pathologies:23	English	Parkinson's Disease	Voice Features of dimension 22
PTDS [138]	Control: 0 Pathologies:42	English	Parkinson's Disease	Voice Features of dimension 22
AVPD [139]	Control: 188 Pathological: 178	Arabic	5 types	Vowels, Sentences, and Counting: 0-10

As mentioned before, the speech signals and contemporaneously recorded EGG signals available in the SVD database [134] are used for the proposed method. The data collection scenario adopted by the SVD database can be visualized in Fig. 3.3. EGG signals are captured utilizing a noninvasive EGG device that measures the contact area between the vocal folds with the two electrodes placed in the proximity of the larynx. Low-amplitude and high-frequency currents are passed through the electrodes. The vocal fold masses are good conductors of electricity. Therefore, during the opening and closed cycles of the glottis, there is a variation in impedance. The EGG device captures this impedance variation, resulting in an EGG signal. The speech signals are recorded with the microphone and sampled at the rate of 50 kHz in the SVD database.

The SVD database is a collection of speech and EGG signals of more than 2000 speakers. It contains recordings of 687 control (i.e., healthy) samples. Among them were 428 females and 259 males. A total of 1356 pathological samples are available in this database. Out of them, 727 females and 629 males. The samples are collected in one session with a normal, high, and low pitch. The sustained vowels are recorded for 1-3 seconds. The sampling frequency is $f_s = 50000$ Hz with 16-bit resolution. This database contains 71 different types of pathologies. Among them, the most common are laryngitis and dysphonia. American Speech-Language-Hearing Association (ASHA) also concludes similar findings in a recent report [140]. According to ASHA's findings, the most frequently diagnosed pathologies among the population (in the age

group 19-60) include dysphonia (20.5%) and laryngitis (12.5%). The subsequent significant pathology is the vocal fold polyps (12%) [141]. Therefore, these three pathologies are considered in this investigation. A brief description and illustrations (Fig. 3. 4) of these three most common voice pathologies are provided below for this work's completeness. Also, the samples considered in this study are the female sustained vowel sound '/a/' with a high tone in the age group of 15 years and above.

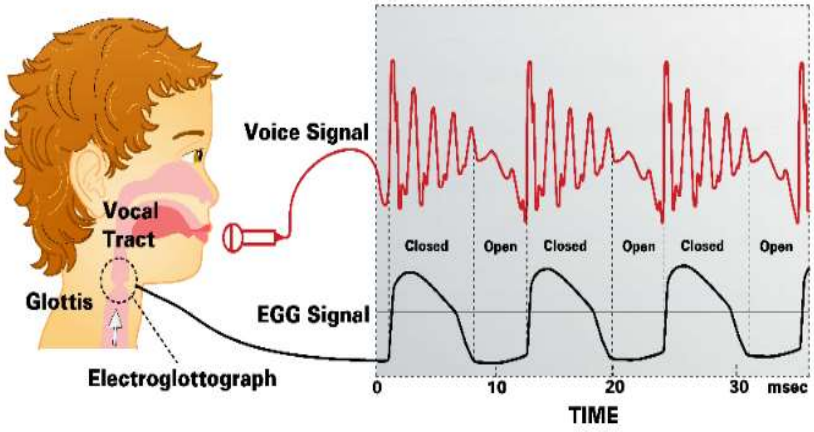


Figure 3.3 The EGG and voice signals [142] as collected in the SVD database.

Laryngitis is an inflammation in the larynx, as shown in Fig. 3.4(a). It is occurred from overuse, smoking, and infection in the larynx [143]. The vocal folds are inflamed and swelled, hence deforming the sounds by obstructing the air. The other reasons for laryngitis are excessive alcohol consumption and GERD [144]. Laryngitis is very common among professional groups, such as singers, actors, telephone operators, lawyers, teachers, referees, coaches, and chemical factory workers. This pathology is also common in children resulting from overused voices. Laryngitis makes the voice sound hoarse and weak. In some cases, it can even make voice undetectable.

Vocal cord polyps are benign masses [145]-[146], just beneath the surface membrane of the vocal cord, as illustrated in Fig. 3.4(b). However, it often results from significant voice use and vocal abuse. It affects the proper vibration of the vocal fold and hence, the quality of the voice. Even a single episode of eventual yelling can cause vocal cord polyps. It can occur on either one or both vocal cords. They have immense blood vessels and hence appear reddish and massive morphologically.

Dysphonia refers to having an abnormal voice [147]-[148] quality. Dysphonia can change the voice suddenly or gradually over time. The voice can be described as hoarse, rough, cracked, weak, breathy, and gravely. Voice may even be completely lost for a while. Dysphonia causes changes in the pitch level. The dysphonic patients complain of pain while speaking, singing, and projecting their voices. Most commonly, dysphonia is caused by an abnormality with the vocal cords. It can also be caused by obstruction of airflow from the lungs. Sometimes, it can result from structural abnormalities of the throat near the vocal cords. It is often a symptom of problems in the vocal folds caused by an upper respiratory infection, a cold, and allergies. Some common types of dysphonia include muscle tension dysphonia, vocal cord paralysis, phono traumatic lesions, recurrent respiratory papillomatosis, paradoxical vocal cord motion, and neurological disorder. Muscle tension dysphonia is the most common one. The cause is abnormal activation of muscle tension. The vocal folds appear long and stretchy, as shown in Fig. 3.4(c). Eventually, that results from the difficulty in phonation, breathing, and swallowing.

The samples of the EGG and corresponding speech signals for the control voice and pathological voices, investigated in this work, are shown in Fig. 3.5(a) and Fig. 3.5(b), respectively. Ideally, the vowel samples for the EGG signals should produce periodic pulses [2].

As shown in Fig. 3.5(a), this is true for the control voice. It is also observed that the open phase, closed phase, and return phase for the EGG samples vary widely depending on the pathologies. Also, the amplitudes of the pathological EGG signals vary significantly compared to the control sample. The vocal cord polyps significantly alter traditional EGG characteristics for the voiced sound of the vowel '/a/'. Like EGG signals, the amplitudes of the speech signals also vary widely, depending on the pathologies, as shown in Fig. 3.5(b). The vocal tract configurations can vary depending on the pathological conditions. The speech samples of the corresponding EGG signals have been reshaped due to different vocal tract configurations.

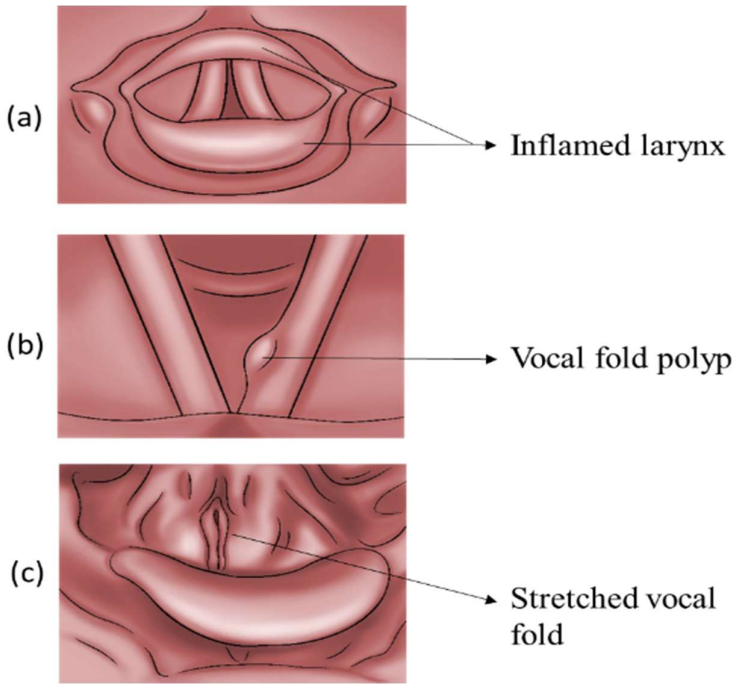


Figure 3.4 The three most common voice pathologies are (a) Inflamed larynx due to laryngitis [149], (b) Vocal cord polyps [145], and (c) Muscle tension dysphonia [150].

3.3 The Proposed Method

In this work, two CNNs are adopted to implement the system. The EGG and speech signals for 150 control and 65 pathological samples are used to train and test CNN-1 separately. As mentioned before, the 65 pathological samples contain a mixture of three different pathologies. The decision made by the CNN-1 is binary (i.e., control/healthy or pathology). Once CNN-1 is adequately trained, with high accuracy, the network is tested with another set of 150 control and 65 pathological samples. After CNN-1 sorts out the pathological voices, these data samples are applied to CNN-2 for pathology classification. The CNN-2 is trained with 64 samples of three pathologies, as mentioned before. Once CNN-2 is trained, another set of 64 samples for the three different pathologies is used to test CNN-2. Finally, the CNN-2 classifies the pathological samples into one of the three pathology types: laryngitis, vocal fold polyps, and dysphonia. The complete flow chart of the proposed system is shown in Fig. 3.6.

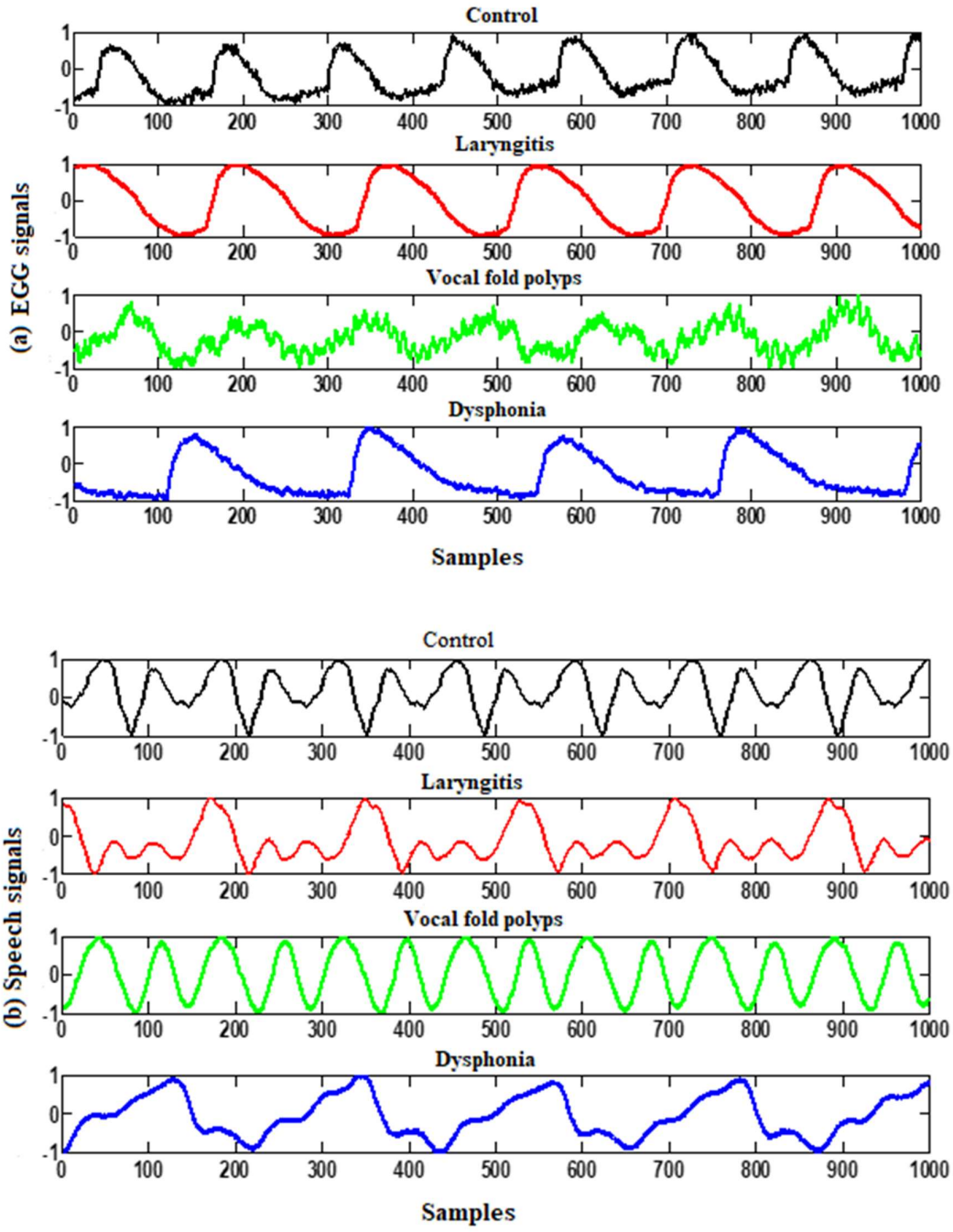


Figure 3.5 The control and pathological samples in the time domain: (a) EGG signals, and (b) speech signals.

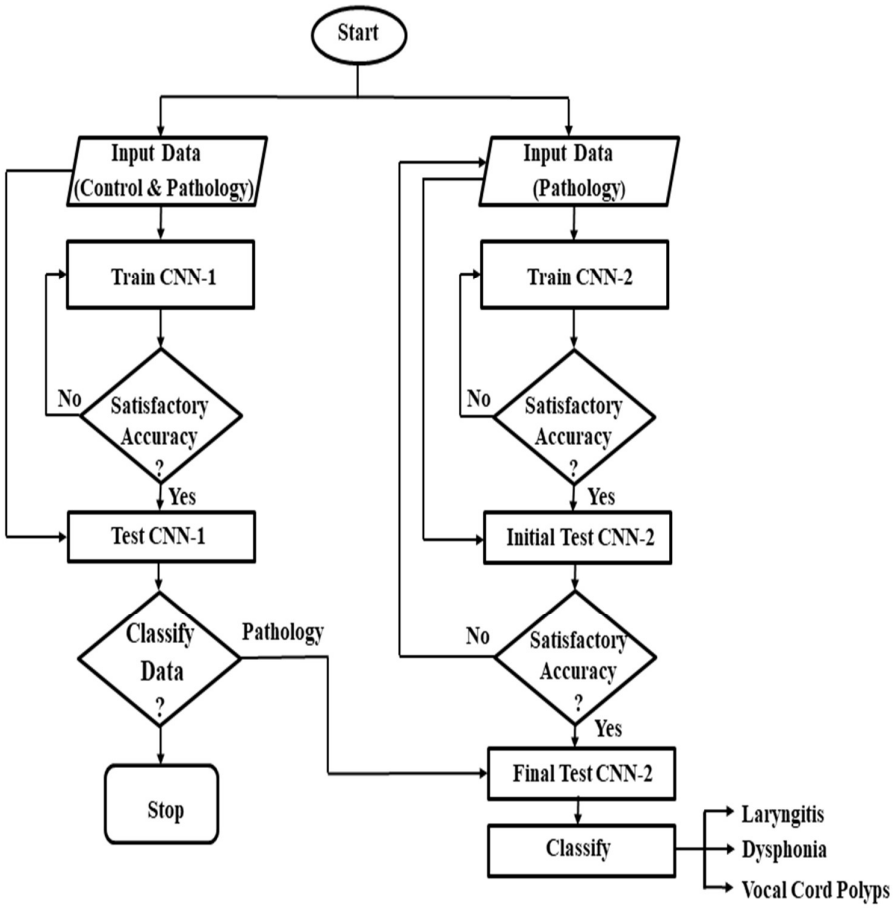


Figure 3.6 The flowchart of the proposed voice pathology detection system.

This work uses the CNN model that originated from a pioneer work published in [151]. The CNN model includes two networks, namely the feature extraction network and classifier network [152]. The input data enters the feature extraction network to produce a feature map based on the input data.

The raw temporal data are fed to the CNN. The primary purpose of this work is to minimize the computation burden on the system. Hence, no audio features are extracted from the speech and EGG samples. The proposed method depends on the built-in feature extraction network of CNN. Since the SVD database provides data of varying lengths for vowel '/a/' sounds, the first $100 \times 100 = 10000$ samples are arbitrarily selected in this work. However, the CNN model used in this work accepts only two-dimensional data as the input. Hence, we reshaped the data and converted

them into a 100×100 matrix to satisfy this requirement. This conversion of the data is illustrated in Algorithm 1.

The feature extraction network consists of a special kind of neural network, of which the synaptic weights are determined via the training process. The feature extractor network consists of piles of convolutional layer and pooling layer pairs. It is widely accepted that a pattern recognition algorithm performs better when the feature extractor network contains more layers (i.e., a deeper network). However, a deeper network is always challenging to train [153]. Considering this, only one convolutional layer is used in the feature extraction network. The convolution layer operates in a very different way compared to other neural network layers. This layer does not employ connection weights and a weighted sum. Instead, it contains filters that convert the input data to produce the feature map. This work uses 20 convolutional filters of size (9×9), as listed in Table 3.2. The feature map that the convolution filters generate is processed through the activation function before the layer yields the output. The ReLU has been used as the activation function in the proposed algorithm. Several other activation functions, including *sigmoid*, and *tanh* have also been investigated in this work. Among these activation functions, the ReLU is selected as this investigation's primary objective is to perform classification tasks without providing a significant computation burden on the system. The main advantage of using the ReLU function is that it does not activate all the neurons simultaneously. Hence, it is far more computationally efficient than the *sigmoid* and *tanh* function.

Table 3.2 The parameters used in CNN Model

Layer	Remarks	Activation Function
Input	(100×100)	---
Convolution	20 convolution filters (9×9)	ReLU
Pooling	1 mean pooling (2×2)	---
Hidden	100 nodes	ReLU
Output	1 node (pathology detection), 3 nodes (pathology classification)	Softmax

Algorithm 1 Data processing using backpropagation algorithm

/* read the data samples of control and pathological voice samples */
1: Read the controlled voice samples, $\overrightarrow{\mathbf{X_c}}$
2: Read the pathological voice samples, $\overrightarrow{\mathbf{X_p}}$
3: Read the training labels, $\vec{\mathbf{D}}$
4: Set random numbers for the convolutional filter weight matrix, $\overrightarrow{\mathbf{W_1}}$
5: Set random numbers for the pooling-hidden layer weight matrix, $\overrightarrow{\mathbf{W_5}}$
6: Set random numbers for the hidden-output layer weight matrix, $\overrightarrow{\mathbf{W_0}}$
/* make all the data of the same length */
7: $\overrightarrow{\mathbf{X_c}} = \overrightarrow{\mathbf{X_c}}[\mathbf{1}:\mathbf{10000}]$;
8: $\overrightarrow{\mathbf{X_p}} = \overrightarrow{\mathbf{X_p}}[\mathbf{1}:\mathbf{10000}]$;
/* convert the data into a two-dimensional array
and make them ready for the input to the convolutional layer*/
9: $\overrightarrow{\mathbf{x_c}} = \mathbf{reshape}(\overrightarrow{\mathbf{X_c}}[\mathbf{100},\mathbf{100}])$;
10: $\overrightarrow{x_p} = \mathbf{reshape}(\overrightarrow{X_{p,}}[\mathbf{100},\mathbf{100}])$;
/* combine the data of control and pathological and form a vector
11: $\vec{\mathbf{X}} = [\,\overrightarrow{\mathbf{x_c}}\ \overrightarrow{\mathbf{x_p}}\,]$
/* initialize the learning rate, momentum factor, and batch size */
12: $\boldsymbol{\alpha} \leftarrow \mathbf{0.01}$;
13: $\boldsymbol{\beta} \leftarrow \mathbf{0.95}$;
14: $\mathbf{batch} \leftarrow \mathbf{2}$;
15: $\mathbf{epoch} \leftarrow \mathbf{30}$;
16: Determine the data size, $\mathbf{N} = \mathbf{length}(\vec{\mathbf{D}})$;
17: Determine the batch size list $\mathbf{blist} = \mathbf{1}:\mathbf{batch}:(\mathbf{N} - \mathbf{batch} + \mathbf{1})$;
18: Determine the number of batches, $\mathbf{M} = \mathbf{length}(\mathbf{blist})$;
19: while (i < epoch) do
20: while (n <M) do
21: / * learning rule */
22: Input data matrix, $\vec{\mathbf{X}}$ % dimension 100 x 100
23: Compute convolution, $\vec{\mathbf{Y}}_{\mathbf{1}} = \overrightarrow{\mathbf{X_1}} * \overrightarrow{\mathbf{W_1}}$; % dimension 92 x 92 x 30
24: Calculate the output of the convolution layer, $\overrightarrow{\mathbf{Y_2}} = \mathbf{ReLU}(\overrightarrow{\mathbf{Y_1}})$;
25: Determine the output $\overrightarrow{\mathbf{Y_3}}$ by 2 x 2 mean pooling % dimension 46 x 46 x 30
26: Reshape the output, $\overrightarrow{\mathbf{Y_4}} = \mathbf{reshape}(\overrightarrow{\mathbf{Y_3}})$; % dimension 63480 x 1
27: Calculate the output of the hidden layer $\overrightarrow{\mathbf{Y_5}} = \mathbf{ReLU}(\overrightarrow{\mathbf{Y_4}}\ \overrightarrow{\mathbf{W_5}})$; % dimension 150 x 1
28: Calculate the final output $\vec{\mathbf{Y}} = \mathbf{Softmax}(\overrightarrow{\mathbf{Y_5}}\ \overrightarrow{\mathbf{W_0}})$;
29: Determine the error at the output-pooling layer, $\vec{\mathbf{e}} = \vec{\mathbf{D}} - \vec{\mathbf{Y}}$ and backpropagate the error
30: Update the weight matrices $\overrightarrow{\mathbf{W_1}}$, $\overrightarrow{\mathbf{W_5}}$, and $\overrightarrow{\mathbf{W_0}}$ using $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$
31: end while
32: end while
33: return $\overrightarrow{\mathbf{W_1}}$, $\overrightarrow{\mathbf{W_5}}$, $\overrightarrow{\mathbf{W_0}}$

The feature map produced by the feature extraction network accentuates the unique features of the original data. Then, the extracted feature map is applied to the classification neural network. The classification neural network operates on the feature map and performs the classification task. The feature map is then passed through the pooling layer. The pooling layer reduces the data size as it combines neighboring data

of a particular area into a single representative value. In this work, a (2×2) matrix is used for pooling the mean value from the input feature maps. The data produced by the pooling layer enters the classifier network, which consists of a hidden layer and an output layer. The classifier network uses a backpropagation algorithm for determining the weight vectors. The hidden layer has 100 nodes that also use the ReLU activation function. The ReLU, as the activation function in the hidden layer, also overcomes the problem of vanishing gradient. The output layers of CNN-1 and CNN-2 contain a different number of output nodes as their purposes are different. Since we use CNN-1 as a binary classifier, the output layer of CNN-1 is constructed with a single node. On the other hand, the output layer of the CNN-2 is built with three nodes as there are three pathologies to classify. The network architectures of the proposed system are shown in Fig. 3.7. The training parameters of the model are set as follows: learning rate is 0.01, momentum is 0.95, and batch size is 2. The other parameters used in the model are listed in Table 3.2. MATLAB 2020 simulation software is adopted to implement the proposed algorithm.

3.4 Simulation Results and Discussion

In voice pathology detection, we need to provide a clinical or diagnostic interpretation of rule-based decisions made with the data samples. Classification decisions are made in the context of medical diagnosis that goes beyond statistical measures of accuracy and validity system [153]. To investigate the performance of our proposed system, we use the following terminologies and performance parameters.

Let us assume that A is the event that a subject has the voice pathology and N is the event that the subject does not have the pathology. If T^+ represents a positive screening test (indicative of the presence of the pathology) and T^- represents a negative screening test (absence of the pathology). The following possibilities will arise [154]-[155]:

True positive (TP) is the situation when the test is positive for the subject with the pathology. Let $P(X)$ indicates the probability of an event, X, the true-positive fraction (TPF) or Sensitivity/Recall is denoted by S^+ and it is given by $P(T^+|A)$ or

$$S^+ = \frac{\text{number of TP decision}}{\text{number of the actual pathological subjects}} \tag{3.2}$$

True negative (TN) represents the case when the test is negative for a subject who does not have the pathology. The true-negative fraction (TNF) or specificity is denoted by S^- and it is given by $P(T^-|N)$ or

$$S^- = \frac{\text{number of TN decision}}{\text{number of the actual healthy subjects}}. \quad (3.3)$$

False-negative (FN) is said to occur when the test is negative for a subject who has the pathology of concern. The probability of this error, known as the false-negative fraction (FNF) and is given by $P(T^-|A)$ or

$$FNF = \frac{\text{number of FN decision}}{\text{TP+number of FN decision}}. \quad (3.4)$$

False-positive (FP) is defined as the case where the result of the test is positive when the individual being tested does not have the pathology. The probability of this type of error or a false alarm, known as the false-positive fraction (FPF) and is given by $(T^+|N)$ or

$$FPF = \frac{\text{number of FP decision}}{\text{TN+number of FP decision}}. \quad (3.5)$$

Accuracy is the most intuitive performance measure, and it is simply a ratio of the correctly predicted observations to the total observations. The accuracy is defined by,

$$\text{Accuracy} = (TP + TN)/(TP + FP + FN + TN). \quad (3.6)$$

Precision or PPV (positive predictive value) is the ratio of correctly predicted positive observations to the total predicted positive observations. The precision is defined by,

$$\text{Precision} = TP/(TP + FP). \quad (3.7)$$

F1 Score is the weighted average of the precision and recall. Therefore, this score takes both FP and FN into account. The F1 Score is defined by,

$$\text{F1 Score} = (2 * \text{Recall} * \text{Precision})/(\text{Recall} + \text{Precision}). \quad (3.8)$$

Negative predictive value (NPV) represents the percentage of the cases labeled as truly negative. The NPV is defined by

$$NPV = TN/(TN + FN). \quad (3.9)$$

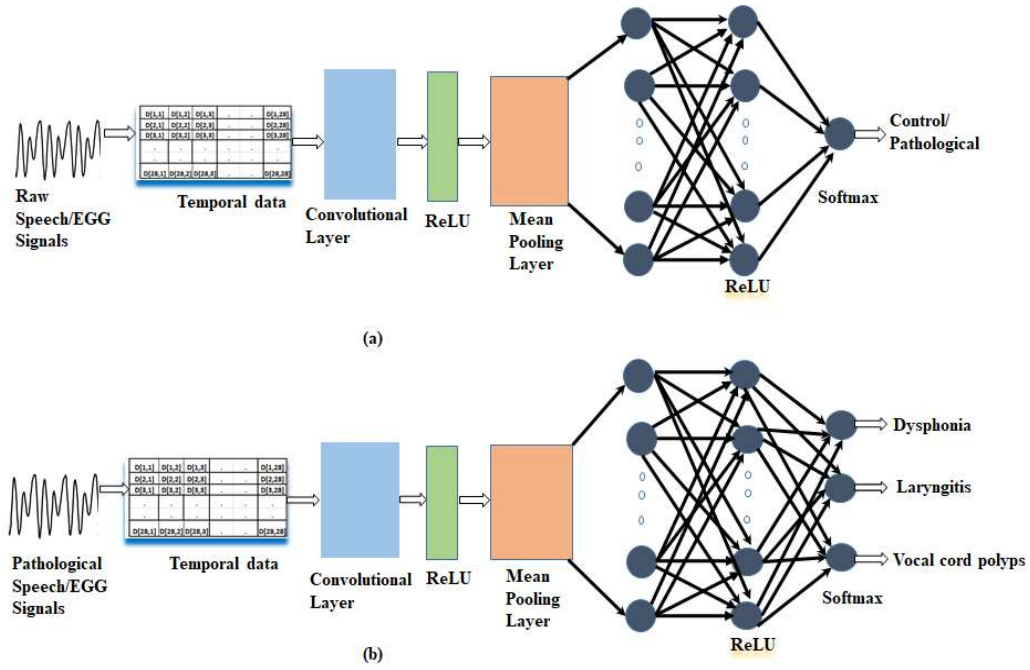


Figure 3.7 The network architecture for (a) binary classifications, and (b) multiclass classifications.

Ten simulations were performed for training and testing; the results of CNN-1 are listed in Table 3.3. This table shows an average training accuracy of 100%, and the testing accuracy of 73.33% achieved with the EGG signals. The performances of the CNN-1, in terms of TP, TN, FP, and FN, are listed in Table 3.4, and the corresponding entries for the classification matrix are provided in Table 3.5. Based on the data presented in Table 3.4 and Table 3.5, we can conclude that CNN-1 can correctly detect voice pathology with an accuracy of 78.34%. On the other hand, it can detect a controlled voice with an accuracy of 65.84%.

Repeating the same experiment with the speech signals, CNN-1 achieves 100.00% and 82.33% accuracy for training and testing, respectively, as listed in Table 3.6. Comparing these values with those presented in Table 3.3, it can be concluded that the speech signals help to achieve higher accuracy (around 9% more) compared to the EGG signals for binary classification.

The system performances in terms of TP, FP, TN, and FN are listed in Table 3.7, and the corresponding entries of the classification matrix are provided in Table 3.8. The data presented in these tables show that the proposed system can detect voice pathology with an accuracy of 90.55% with speech signals, which is 12.21% more than the case with EGG signals, as listed in Table 3.5. Comparing the data presented in Table

3.8 to that in Table 3.5, we can also conclude that the proposed system is less prone to misclassify pathological subjects as control subjects when the speech signal is used instead of the EGG signal. The statistical performance comparisons of the proposed system with the EGG and speech signals are listed in Table 3.9. This table shows that the performance of CNN-1 with speech signals outperforms the one with EGG signals in terms of accuracy, precision, recall, F1 Score, and NPV. These scores are almost 6-13% higher in speech signals than their EGG signals counterparts.

The mixed pathological EGG samples of laryngitis, dysphonia, and vocal cord polyps are used to train the CNN-2. The data presented in Table 3.10 shows that CNN-2 achieves 100% and 88.67% accuracy in training and testing, respectively. After successful training and testing of CNN-2 with the pathological data, the pathological EGG samples discriminated by CNN-1 are applied as the input of CNN-2 as depicted in the flowchart presented in Fig. 3.6. The simulation results of the final classification performed by the CNN-2 are listed in Table 3.11. The results show that CNN-2 detects dysphonic pathology more accurately than laryngitis and vocal cord polyps. The classification matrix presented in Table 3.12 shows that the system has decided 16.5% of the data as vocal cord polyp and 6.5% as dysphonic although the original data belong to laryngitis pathology. The CNN-2 misclassifies vocal cord polyp as laryngitis with a probability of 21.83%. However, it misclassifies the dysphonic pathology as polyps with a probability of 20%.

Similar experiments have been repeated for the mixed speech pathological samples of laryngitis, dysphonia, and vocal cord polyps. The results are presented in Table 3.13-3.15. The data presented in Table 3.13 shows that CNN-2 achieves an accuracy of 100% and 76.48% in training and testing, respectively. After successful training and testing of CNN-2 with the pathological data, the pathological speech samples discriminated by CNN-1 are applied as the input of CNN-2 as depicted in the flowchart presented in Fig. 3.6. The simulation results of the final classification performed by the CNN-2 are listed in Table 3.14. The results show that CNN-2 can detect dysphonic pathology more accurately (83.33%) compared to laryngitis (78.83%) and vocal cord polyps (63.00%). The classification matrix presented in Table 3.15 shows that the system has decided 21.16% of the data as vocal cord polyp, although the original data belong to laryngitis pathology. The CNN-2 misclassifies vocal cord polyp as laryngitis with a probability of 29.67%. The CNN-2 also misclassifies dysphonic as laryngitis, with a possibility of 17.66%.

The comparison of the performance parameters in terms of precision/PPV, recall/S+, and F1 Score for CNN-2 with EGG and the speech signals are listed in Table 3.16. The results show that the performance parameters are always higher with the dysphonic voice compared to the other two pathology types. However, the uneven confusion matrix compelled us to focus on the F1-Scores where laryngitis and vocal fold polyps achieve better results with EGG signals (0.775 and 0.731, respectively) as compared with speech signals (0.697 and 0.684, respectively). We can also conclude that the dysphonic voice is easier to detect compared to laryngitis and vocal cord polyps using either speech or EGG signals (F1 Scores are 0.868 and 0.858, respectively). The experimental results manifest that control and mixed pathological speech samples are better candidates for binary classification than EGG signals. The multi-classification performance is superior for EGG signals and strongly depends on the pathology attributes. The EGG signals that identify the phonation activity of the vocal folds resemble a better prognosis for vocal cord polyp and laryngitis. Clinically, the strong correlation of these two voice pathologies with the vocal fold justifies the result. The diseases in the vocal fold or vocal tract should have a direct impact on the voice. On the practical side, the structural presence of vocal fold polyps interrupts the glottal closure. The size and stiffness of polyps directly interfere with the vocal fold's vibratory pattern, resulting in more hoarseness. While large polyps tend to induce sub-harmonics and chaos, small polyps may not influence much the periodicity of vocal fold vibrations. Hence, the variable morphology of vocal fold polyps makes it harder for the system to identify them compared to the other two pathologies using EGG or speech signal as demonstrated in this study. The dysphonic voice identification is almost equally efficient using either EGG or speech signals.

Finally, the performances of the proposed algorithm are compared with other related works available in the literature and the comparison is listed in Table 3.17. The data presented in the table shows that the proposed algorithm outperforms the works published in [156]. The performances of the proposed algorithm are also comparable with that of the algorithms listed in Table 3.17. However, the algorithms presented in [7], [21], and [157] provided much higher accuracy (92.8%, 99.69%, and 99.8%, respectively) than the proposed algorithm. The algorithm presented in [21] needs to compute the Mel-spectrogram from the speech signals. It also necessitates the computation of the closed and open quotient, peak amplitude, peak width, and cepstral features from the EGG signals. Also, the algorithm presented in [7] requires computing

the autocorrelation and entropy of the voice signals. It also entails a searching algorithm to locate the peak values and their lags in the autocorrelation function. The work in [157] demands the computation of correlation functions from voice signals. The work presented in [158] required a significant number of computations to extract the features from the voice samples. The proposed algorithm avoids those kinds of computation burdens on the system as the original/raw temporal speech and EGG samples are directly used for voice pathology detection. Another major advantage of the proposed system is that it deals with fewer neural network parameters than other systems. For example, the algorithm presented in [159] uses a pre-trained CNN network called VGG16, which contains 13 convolutional layers and 138 million associated parameters. The algorithm presented in [158] uses a combination of networks (i.e., CNN+MLP and CNN+LSTM), and hence the computation burden is also significant. On the other hand, the proposed system uses only one convolutional layer, two hidden layers, and one-pooling layer (see Fig. 3.7). Based on the above-mentioned observation, we can conclude that the proposed system provides less computation burden, achieving accuracy comparable to some other related works.

3.5 Conclusion

An automated noninvasive pathological voice detection system using both the EGG and the speech signals has been presented in this Chapter. It has been always an argument about using the EGG signals or the speech signals for a pathological voice detection system. This Chapter provided a rigorous in-depth etiology of these two signals. The sources of origin of these signals and their characteristics revealed their performance for binary and multi-classification of pathological voices. The better detection accuracy (80.30%) for binary classification has been achieved with the raw temporal speech signals. But the performance for multi-classification has exhibited better F1 Scores for laryngitis (0.775) and vocal fold polyps (0.731) with EGG signals than speech signals. The best F1 Scores were achieved for classifying dysphonia with speech (0.868) and EGG signals (0.858).

This work has also shown that a small dataset would be enough to train the proposed dual CNN algorithm with high accuracy, relying on its in-built feature extractor network. Another advantage of the proposed system is that it does not need a separate feature extraction network. Also, the system can extract discriminative features from the raw samples as opposed to conventional algorithms. Hence, the proposed

system is much faster than most other feature-based systems requiring a special type of processor to overcome the computation burden.

TABLE 3.3
TRAINING AND TESTING ACCURACIES OF CNN-1 WITH THE EGG SIGNALS ((BINARY CLASSIFICATION)

Accuracy of the EGG signals		
Simulation No.	Training (%)	Testing (%)
1	100.00	73.33
2	100.00	73.33
3	100.00	73.33
4	100.00	73.33
5	100.00	73.33
6	100.00	73.33
7	100.00	73.33
8	100.00	73.33
9	100.00	73.33
10	100.00	73.33
Average	100.00	73.33
Std. Dev.		0.00

TABLE 3.4
THE PERFORMANCES OF CNN-1 WITH THE EGG SIGNALS (BINARY CLASSIFICATION)

Simulation. No.	*TP* (%)	*TN* (%)	*FP* (%)	*FN* (%)
1	77.78	66.67	33.33	22.22
2	77.78	66.67	33.33	22.22
3	83.33	58.33	41.67	16.67
4	77.78	66.67	33.33	22.22
5	77.78	66.67	33.33	22.22
6	77.78	66.67	33.33	22.22
7	77.78	66.67	33.33	22.22
8	77.78	66.67	33.33	22.22
9	77.78	66.67	33.33	22.22
10	77.78	66.67	33.33	22.22
Average	78.34	65.84	34.16	21.67

TABLE 3.5
THE CONFUSION MATRIX OF CNN-1 WITH THE EGG SIGNALS (BINARY CLASSIFICATION)

	Prediction (%)	
Actual	Control	Pathology
Control	65.84% (S^-)	34.16% (*FPF*)
Pathology	21.67% (*FNF*)	78.34% (S^+)

TABLE 3.6
TRAINING AND TESTING ACCURACIES OF CNN1 WITH THE SPEECH SIGNALS (BINARY CLASSIFICATION)

Accuracy of the speech signals		
Simulation No.	Training (%)	Testing (%)
1	100.00	90.00
2	100.00	86.67
3	100.00	90.00
4	100.00	80.00
5	100.00	76.67
6	100.00	76.67
7	100.00	83.34
8	100.00	76.67
9	100.00	86.67
10	100.00	76.67
Average	100.00	82.34
Std. Dev.		5.383

TABLE 3.7
THE PERFORMANCES OF CNN-1 WITH THE SPEECH SIGNALS (BINARY CLASSIFICATION)

Simulation No.	*TP* (%)	*TN* (%)	*FP* (%)	*FN* (%)
1	100.00	75.00	25.00	0.00
2	94.44	75.00	25.00	5.56
3	94.44	83.33	16.67	5.56
4	77.78	83.33	16.67	22.22
5	83.33	66.67	33.33	16.67
6	83.33	66.67	33.33	16.67
7	94.44	66.67	33.33	5.56
8	88.89	58.33	41.67	11.11
9	94.44	75.00	25.00	5.56
10	94.44	50.00	50.00	5.56
Average	90.55	70.00	30.00	9.45

TABLE 3.8
THE CONFUSION MATRIX OF CNN-1 WITH THE SPEECH SIGNALS (BINARY CLASSIFICATION)

	Prediction (%)	
Actual	Control	Pathology
Control	70.00% (S^-)	30.00% (*FPF*)
Pathology	9.45% (*FNF*)	90.55% (S^+)

TABLE 3.9
THE PERFORMANCE COMPARISON OF CNN-1 WITH EGG AND SPEECH SIGNALS (BINARY CLASSIFICATION)

Measures	Speech signal	EGG Signal
Accuracy	**0.803**	0.721
Precision/ PPV	**0.751**	0.696
Recall/Sensitivity	**0.906**	0.783
F1 Score	**0.821**	0.737
NPV	**0.881**	0.752

TABLE 3.10
THE TRAINING AND TESTING ACCURACIES OF CNN-2 WITH THE EGG SIGNALS (MULTI-CLASSIFICATION)

Accuracy of the EGG signals (%)		
Simulation No.	Training (%)	Testing (%)
1	100.00	86.67
2	100.00	93.33
3	100.00	86.67
4	100.00	86.67
5	100.00	86.67
6	100.00	86.67
7	100.00	86.67
8	100.00	93.33
9	100.00	86.67
10	100.00	93.33
Average	100.00	88.67
Std. Dev.		3.05

TABLE 3.11
THE TESTING ACCURACIES OF CNN2 WITH THE BINARY CLASSIFIED EGG SIGNALS (MULTI-CLASSIFICATION)

Simulation No.	Laryngitis (%)	Polyp (%)	Dysphonic (%)
1	75.00	80.00	80.00
2	100.00	80.00	80.00
3	60.00	80.00	80.00
4	75.00	80.00	80.00
5	75.00	80.00	80.00
6	75.00	80.00	80.00
7	75.00	80.00	80.00
8	100.00	80.00	80.00
9	75.00	80.00	80.00
10	60.00	66.67	80.00
Average	77.00	78.67	80.00
Std. Dev.	12.88	3.90	0

TABLE 3.12
THE CLASSIFICATION MATRIX FOR CNN-2 WITH THE EGG SIGNALS (MULTI-CLASSIFICATION)

	Prediction (%, per class)		
Actual	Laryngitis	Polyp	Dysphonic
Laryngitis	77	16.5	6.5
Polyp	21.83	78.67	0
Dysphonic	0	20	80

TABLE 3.13
THE TRAINING AND TESTING ACCURACIES OF CNN-2 WITH THE SPEECH SIGNALS (MULTI-CLASSIFICATION)

Accuracy of the speech signals		
Simulation No.	Training (%)	Testing (%)
1	100.00	66.66
2	100.00	83.33
3	100.00	76.47
4	100.00	71.43
5	100.00	80.00
6	100.00	80.00
7	100.00	64.70
8	100.00	83.34
9	100.00	82.35
10	100.00	76.47
Average	100.00	76.48
Std. Dev.		6.43

TABLE 3.14
THE TESTING ACCURACIES OF CNN-2 WITH THE BINARY CLASSIFIED SPEECH SIGNALS (MULTI-CLASSIFICATION)

Simulation No.	Laryngitis (%)	Polyp (%)	Dysphonic (%)
1	66.67	50	83.33
2	83.33	60	83.33
3	83.33	60	83.33
4	75.00	60	80.00
5	100.00	60	80.00
6	100.00	60	80.00
7	50.00	60	83.33
8	80.00	60	83.33
9	83.33	80	83.33
10	66.67	80	83.33
Average	78.83	63	82.33
Std. Dev.	14.40	9.0	1.52

TABLE 3.15
THE CLASSIFICATION MATRIX FOR CNN-2 WITH THE SPEECH SIGNALS (MULTI-CLASSIFICATION)

	Prediction (%, per class)		
Actual	Laryngitis	Polyp	Dysphonic
Laryngitis	78.83	21.16	0
Polyp	29.67	63	7.33
Dysphonic	17.66	0	82.33

TABLE 3.16
COMPARISON OF PERFORMANCE MEASURES FOR CNN-2 ((MULTI-CLASSIFICATION), PER CLASS

Measures	Speech	EGG	Speech	EGG	Speech	EGG
	Laryngitis		Polyps		Dysphonic	
Precision/ PPV	0.625	0.779	0.749	0.683	0.918	0.925
Recall/ S+	0.788	0.770	0.630	0.787	0.823	0.800
F1 Score	**0.697**	**0.775**	0.684	**0.731**	**0.868**	**0.858**

TABLE 3.17
PERFORMANCE COMPARISON WITH OTHER RELATED WORKS

Study	Datasets	Assessment Measures	Features	Classifier	Best Results/Findings
M. Shamim et al. in [21]	SVD	*Accuracy, Sensitivity, Specificity*	Voice signals: Mel-spectrogram EGG signals: Closed quotient, open quotient, peak amplitude, peak width, cepstral feature	GMM	*Accuracy*:(detection) 92.8% (voice signal), 77.7% (EGG signal).
A. Al-Nasheri et al. in [7]	SVD, AVPD, MEEI	*Accuracy*, *Sensitivity*, *Specificity*	Voice signals: Autocorrelation: Peaks, and their lags, Entropy	SVM	*Accuracy*: (MEEI database) 99.69% (detection), 99.54% (classification).
A. Al-Nasheri et al. [157]	SVD, AVPD, MEEI	*Accuracy*, *Sensitivity*, *Specificity*	Voice signals: Correlation: Peaks and their lags	SVM	*Accuracy*: (MEEI database) 99.80% (detection), 99.25% (classification). 1000-8000 Hz is the most significant band for voice pathology detection.
M. Alhussein and G. Muhammad [159]	SVD, MEEI	*Accuracy*, *Sensitivity*, *Specificity*	Voice signals: Octave spectrogram	Pre-trained CNN	*Accuracy*: (SVD database) 97.5% (detection)
S. Kadiri and P. Alku in [156].	Hospital Universitario Príncipe de Asturias (HUPA), SVD	*Accuracy*, *Sensitivity*, *Specificity*, Area under the curve (AUC), Equal Error Rate (EER)	Glottal source waveform: Time and frequency-domain features	SVM	(HUPA database) *Accuracy*: 78.37% (detection) EER: 0.207, AUC: 0.84. The proposed method outperforms while combining glottal features with conventional MFCCs and PLP.
N.P. Narendra and P. Alku in [158]	Universal access speech (UA-Speech), TORGO database, Universidad Politécnica de Madrid (UPM)	*Accuracy*, *Sensitivity*, *Specificity*	Temporal raw glottal flow, Temporal raw speech	CNN+MLP, CNN+LSTM	*Accuracy*: (detection) 87.93%: UA-Speech 81.12%: TORGO 76.66 %: UPM. Better detection accuracy with glottal flow compared to speech samples.
Proposed System	SVD	*Accuracy Sensitivity Precision F1 Score NPV*	Temporal raw EGG, Temporal raw speech	CNN	*Accuracy*: binary classification /detection 80.3% : Speech 72.1%: EGG *F1score*: multi-classification Dysphonia: 0.868 (Speech) Polyps: 0.731 (EGG) Laryngitis: 0.775 (EGG) Better detection accuracy with speech signals. Overall performance for multi-classification is better for EGG signals.

CHAPTER 4
A PATHOLOGICAL VOICE IDENTIFICATION TECHNIQUE THROUGH COCHLEAR IMPLANT PROCESSING SYSTEM

This Chapter presents a pathological voice identification system employing signal processing techniques through cochlear implant models. The fundamentals of the biological process for speech perception are investigated to develop this technique. This work considers two cochlear implant models: one uses a conventional bank of bandpass filters, and the other uses a bank of optimized gammatone filters. The critical center frequencies of those filters are selected to mimic the human cochlear vibration patterns caused by audio signals. The proposed system processes the speech samples and applies a CNN for final pathological voice identification. The results show that the two proposed models adopting bandpass and gammatone filter banks can discriminate the pathological voices from healthy ones, resulting in F1 Scores of 77.6% and 78.7%, respectively, with speech samples. The obtained results of this work are also compared with those of other related published works.

4.1 Introduction

This work investigates the possibility of using the existing technology of the cochlear simulation model noninvasively for detecting pathological voices. The clinical tools used by physicians rely on invasive technology that is unpleasant for the patients. Additionally, they sometimes rely on subjective assessment, especially for voice pathology that lacks structural abnormality. To overcome these limitations, this Chapter addresses a signal processing and deep learning-based technology that can help clinicians with noninvasive objective assessment of voice disorder, thus providing relief for the patients from painful processes and avoiding the misdiagnosis that may result from subjective evaluations.

Many pathological voice detection systems have been published in the literature. However, the novelty of this work is it uses the cochlear simulation model to implement a pathological voice detection system. The voice samples are processed using a cochlear simulation model, and then the processed voice samples are applied to the input of a CNN for final classification.

Cochlear implants are sensory prosthetic devices. They can establish the functional hearing of the listeners with severe hearing loss. This is achieved by establishing direct electrical stimulation to the auditory nerves for people with damaged

hair cells in the basilar membrane. These hair cells are tuned at different frequencies to aid hearing perception for people with no hearing impairment [160]. A typical cochlear implant system includes several signal processing steps: (a) removal of the D.C. component, (b) pre-emphasis, (c) division of the signal into a set of channels, (d) rectification, and (e) lowpass filtering. Among these signal processing steps, the most critical is dividing the signal into several channels using a filter bank. The center frequencies and the bandwidth of these filters are determined based on the human cochlear vibration patterns caused by the audio samples. In this work, two models have been considered for the filter bank. One model uses a bank of bandpass filters, and the other uses gammatone filters. A bank of bandpass filters is commonly used in commercially available cochlear implants. However, recently, researchers are recommending using gammatone filters instead. The main advantages of the gammatone filters are that they (a) provide an appropriate "pseudo-resonant" frequency transfer function, (b) demonstrate a simple impulse response, and (c) support efficient hardware implementation [161]. Finally, the processed audio features are applied to the input of a CNN for classification. The main contributions of this work are as follows:

- It develops a novel, non-invasive pathological voice detection algorithm based on speech signal processing that mimics the biological process of speech perception and a deep learning approach.
- It extracts audio information using gammatone and conventional bandpass filters to examine their efficacy for pathological voice identification.
- It eliminates the necessity of choosing suitable features from speech samples to aid the classification mechanism.
- It achieves a reasonably high classification accuracy without overwhelming the computation burden on the system.
- It provides a detailed performance analysis of the proposed system in terms of accuracy, precision, recall, NPV, F1 Score, and G-mean.
- It compares the performances of the proposed system with other related works to demonstrate its effectiveness.

4.2 Materials and Methods

This investigation collected control (i.e., normal) and laryngitis voice samples from the SVD database [133]. The samples contain the recordings of the following components: (a) vowels '/i/', '/a/', and '/u/' produced at a normal, high, and low pitch, (b) vowels

'/i/', '/a/', and '/u/' with rising and falling pitch, and (c) sentence, "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?"). In this investigation, the sentence, "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?") has been used. The main reason is that the sentence speech samples contain both voiced and unvoiced components. On the other hand, the vowel speech samples contain only the voiced component. Moreover, the sentence speech samples have articulatory and other linguistic confounds that often do not exist with the vowel samples. Figure 4.1 shows the time domain plots for control (i.e., healthy) and laryngitis sentence speech samples randomly collected from the SVD database. It is observed in the figure that the laryngitis voice sample suffers from irregular distortion in both magnitude and shape compared to that of the healthy sample. In addition, the laryngitis voice samples exhibit a more extended unvoiced segment than the vowel samples.

The basic building blocks of the proposed system are shown in Fig. 4.2. This model was derived based on the commercially available Clarion 1.2 processor [162]-[164] introduced by Advanced Bionics Corporation in cooperation with the University of California and the Research Triangle Institute.

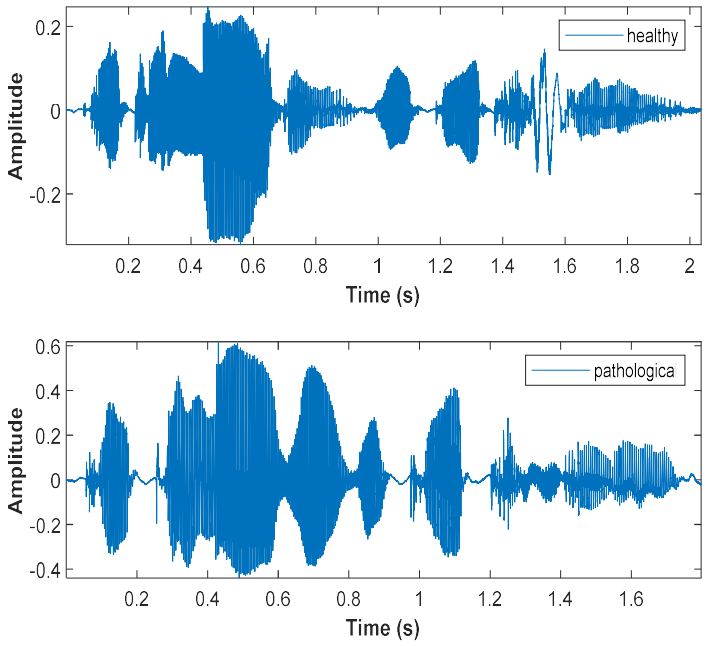


Figure 4.1 The healthy and pathological voice samples of "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?").

As shown in Fig. 4.2, the system model can be broadly classified into three major sub-systems: (a) pre-processing, (b) cochlear modeling, and (c) classification. The pre-processing sub-system consists of three signal processing steps: down-sampling, D.C. removal, and pre-emphasis. In a Clarion processor, the acoustic signal is processed at the rate of 13000 samples/s. On the other hand, the voice samples available in the SVD database have a sampling frequency of 50000 samples/s. Hence, the voice signals were down-sampled to 13000 samples/s using the MATLAB built-in *resample* function. The *resample* function utilizes a built-in anti-aliasing (lowpass) finite impulse response (FIR) filter to minimize the effects of aliasing that occur due to the down sampling operation. Afterward, the D.C. component of the speech signals was removed.

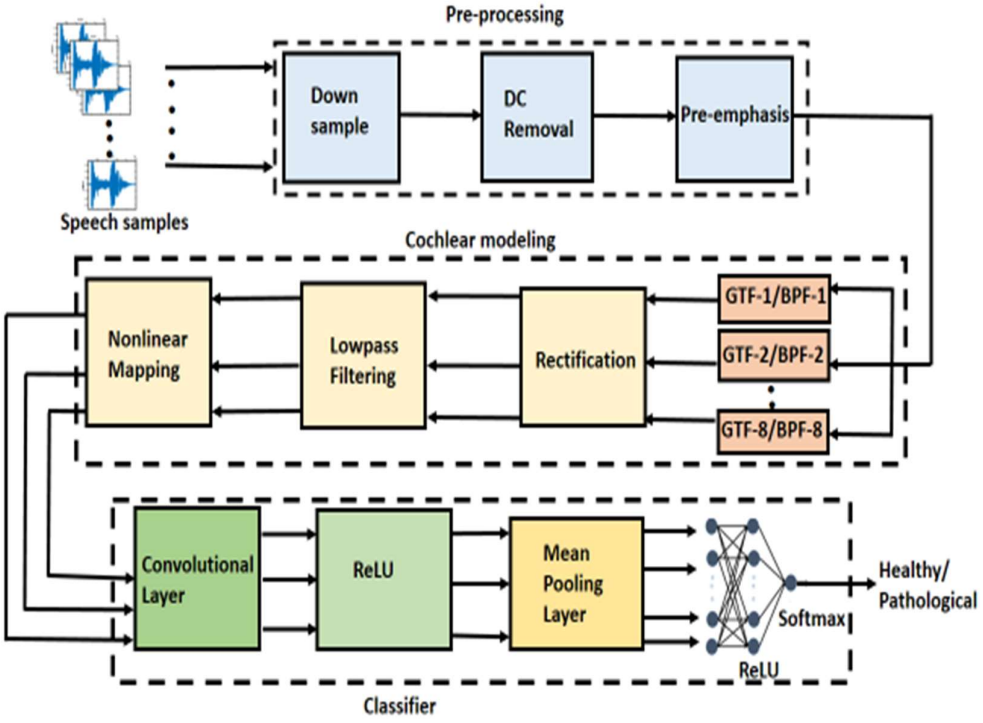


Figure 4.2 The proposed system, comprised of pre-processing, cochlear modeling, and classifier.

Most of the energy in the speech signal is concentrated in the lower frequency components of its spectrum, and generally, the energy drops at a rate of 2.0 dB/kHz [165]. This rapid reduction in energy leads to a problem for further subsequent processing of speech signals. To overcome this limitation, the high-frequency components of the speech signals were boosted by a pre-emphasis filter, which was

designed based on the model presented in [166]. The magnitude response of the pre-emphasis filter is shown in Fig. 4.3, which has a cut-off frequency of 2000 Hz and a roll-off of 3 dB/octave. It compensates for the rapid reduction of the energy in the low-frequency components of the audio signal. Additionally, it better optimizes CPU consumption.

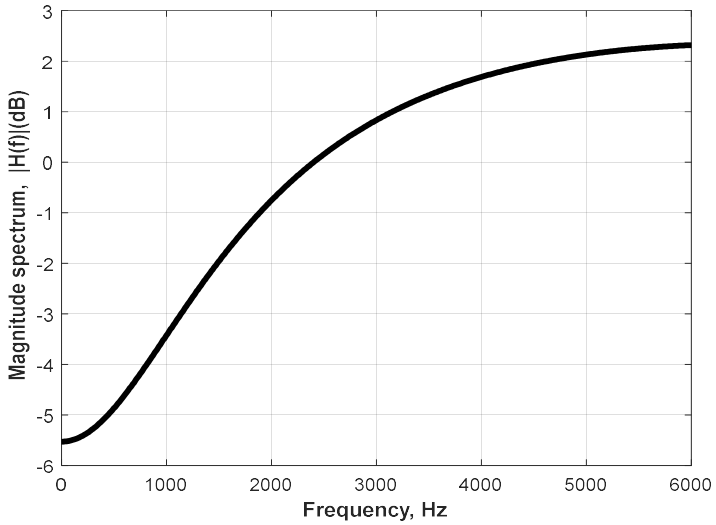


Figure 4.3 The magnitude spectrum of the pre-emphasis filter with a cut-off frequency of 2000 Hz.

It is also shown in Fig. 4.2 that the cochlear modeling sub-system consists of a bandpass filter, rectifier, lowpass filter, and a non-linear mapper. The pre-processed speech signals were divided into eight channels by using eight filters. These filters were designed based on the specifications mentioned in [166]. The center frequency and the bandwidth of these eight filters are listed in Table 4.1. These eight filters were designed by using the third-order Butterworth prototype filters. The table demonstrates that the filters' bandwidth is logarithmically spaced from 265 to 1136 Hz, mimicking the frequency response of the basilar membrane. The basilar membrane is mechanically tuned at different frequencies, and it plays a vital role in distributing sound energy by frequencies along the cochlea's length, as shown in Fig. 4.4. The designed band pass filter has the lowest center frequency is 394 Hz (the center frequency of the first filter), and the highest center frequency is 4871 Hz (the center frequency of the eighth bandpass filter). The magnitude spectrum of these eight bandpass filters is shown in Fig. 4.5.

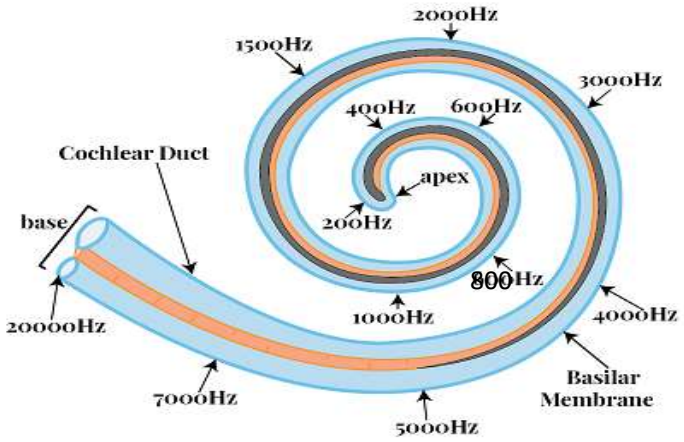


Figure 4.4 The tuning frequencies of the basilar membrane [167].

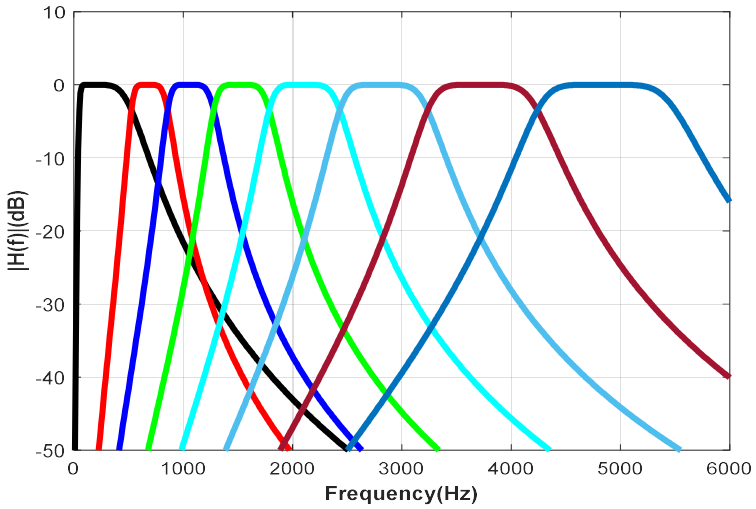


Figure 4.5 The magnitude response of the bandpass filter bank.

The next signal processing steps include envelope detection and lowpass filtering. This work used a full-wave rectifier as an envelope detector, and an eighth-order FIR filter was used as a lowpass filter. This lowpass filter was designed by using the Hamming window function. Several window functions, namely Hanning, Blackman, Bartlett, and Hamming, have been investigated in this work. The main advantages of these window functions are that they taper at their ends and avoid unnatural discontinuity in the speech segment. They also minimize the distortion in the underlying spectrum. Finally, the Hamming window function was selected as it provided the minimum passband ripple and maximum stopband attenuation compared to the other investigated window functions [168].

Table 4.1 The bandwidth and center frequencies of the eight filters

Bandwidth, Hz	Center Frequency, Hz
265	394
331	692
431	1064
516	1528
645	2109
805	2834
1006	3740
1136	4871

Finally, the detected signal envelope in each channel was used to modulate a biphasic pulse train. A non-linear mapping technique was used to produce the biphasic pulse train so that the interferences of the pulses in different channels were minimized.

The eight filters (mentioned above) were replaced by eight gammatone filters in the second model while using the same other components. The pre-processed audio signals were divided into eight channels by using these eight gammatone filters. The name gammatone comes from the fact that the envelope of the impulse response of those filters resembles the gamma function. Moreover, the fine structure of the impulse response is a tone at the center frequency of the filter, f_0 [169]-[170]. Those gammatone filters perform spectral analysis and convert an acoustic wave into a multichannel representation by mimicking the basilar membrane motion [171]. The gammatone filter has an impulse response that is like that of a cat's cochlea [172], and it is defined by:

$$h(t) = ct^n e^{-2\pi} \quad cos(2\pi f_0 t + \varphi)u(t), \tag{4.1}$$

where c is a constant, n is the filter order, b is the temporal decay coefficient, f_0 is the center frequency of the filter, φ is the carrier phase, and $u(t)$ is the unit step function. The filter order, n, controls the relative shape of the envelope that becomes less skewed when n increases. The carrier phase, φ, determines the relative position of the envelope. Let us assume that the carrier component is denoted by $s(t) = cos(2\pi f_0 t + \varphi)$ and the gammatone distribution function is defined by $r(t) = t^{n-1}e^{-2\pi bt}u(t)$. Hence, the impulse response of the gammatone filter can be expressed as $h(t) = cs(t)r(t)$. The parameter b determines the duration of the impulse response and hence, determines the

bandwidth of the gammatone filters, and the parameter n determines the tuning or quality factor (Q) of the filter. Fig. 4.6 shows the impulse response of the gammatone filter with its constituent components. In the plot, the factor c was set to $\frac{b^n}{(n-1)!}$ to make the area under the curve of gamma distribution equal to one [161]. The temporal decay coefficient b was set to 125, and the carrier frequency, f_0, was chosen to be 1000 Hz. The shape of the magnitude characteristic of the gammatone filters with order 4 is very similar to that of the $roex$ function [173] that is commonly used to represent the magnitude response of the human auditory filter [174]-[175]. The Fourier transform of the $h(t)$ is given by $H(f)$ and it can be expressed as

$$H(f) = \frac{c}{2}(n-1)!\,(2\pi b)^{-n}\left[e^{j\varphi}\left(1+j\frac{(f-f_0)}{b}\right)^{-n}\right] + \frac{c}{2}(n-1)!\,(2\pi b)^{-n}\left[e^{-j\varphi}\left(1+j\frac{(f+f_0)}{b}\right)^{-n}\right] \quad (4.2)$$

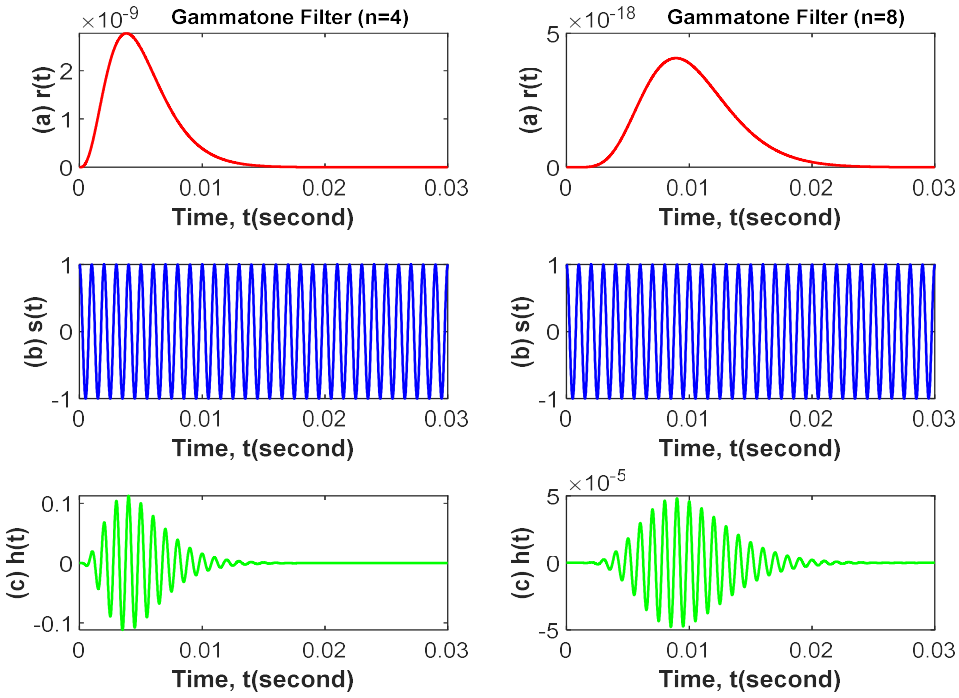


Figure 4.6 The components of a gammatone filter: (a) gammatone distribution function, (b) the carrier tone, and (c) impulse response.

A complete derivation of the $H(f)$ can be found in Appendix A. The impulse response, $h(t)$, and the transfer function, $H(f)$, of the gammatone filter with varying f_0/b are plotted in Fig. 4.7, which shows that the two frequency components of the gammatone filters do not interfere with each other when $f_0/b > 8$. In this work, we selected $f_0/b = 9$. Another advantage of selecting $f_0/b = 9$ is that the bandwidth

becomes proportional to b, and it is claimed in [176] that the bandwidth (equivalent rectangular bandwidth) becomes independent of f_0 when $f_0/b > 3$. The detailed proof is shown in Appendix B. The center frequency and the bandwidth of the gammatone filters are listed in Table 4.2, while the magnitude spectrum of the gammatone filter bank is shown in Fig. 4.8. The filters are logarithmically spaced in frequency resolution that is similar to the basilar membrane's motion, as shown in this figure 4.4.

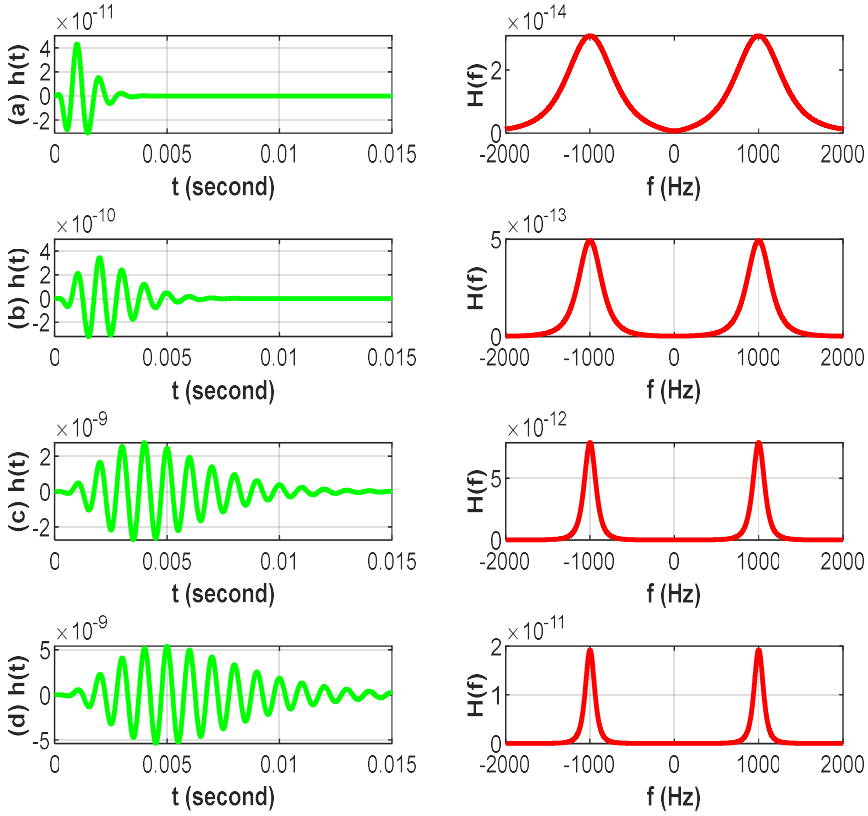


Figure 4.7 The filter impulse responses, $h(t)$, and their corresponding spectrums, $H(f)$, for: (a) $f_0/b = 2$, (b) $f_0/b = 4$, (c) $f_0/b = 8$, and (d) $f_0/b = 10$.

Another main system component is the classifier, as shown in the proposed system's last sub-system presented in Fig. 4.2. The processed signal from the cochlear model is applied to the input of a classifier for binary classification. In this work, a CNN was employed for this purpose. The CNN presented in [152] was adopted and optimized to implement the proposed system. The CNN includes feature extraction and classifier networks. The feature extractor network produces a feature map based on the input data. The feature map accentuates the unique features from the original data.

Consequently, the extracted feature map was applied to the classification neural network. The classification neural network operates on the feature map and performs classification functions. The feature extractor network consists of a special kind of neural network, of which the synaptic weights are determined via the training process. Usually, the feature extractor network consists of piles of convolutional layer and pooling layer pairs, as shown in Fig. 4.2. It is widely accepted that pattern recognition algorithms perform better when the feature extractor network contains more layers. However, it is always challenging to train them as it incurs a substantial computational burden on the system [153]. Considering this limitation, this work used one convolutional layer as a feature extractor network.

Unlike other conventional neural networks, no connection weights or a weighted sum are employed in the convolutional layer. Instead, filters are used to convert the input data to produce a feature map. In this work, 20 convolutional filters of size 11×11 were used. An activation function processes the feature map produced by the convolutional filters. In this work, the ReLU is used as the activation function. The output produced by the convolutional layer is then passed through the pooling layer. The pooling layer reduces the data size by combining the neighboring data of a certain area into a single representative value. In this work, a 2×2 matrix was used for pooling the mean value from the input data. The data produced by the pooling layer enters the classifier network, which consists of a hidden layer and an output layer. A backpropagation algorithm was used to determine the weight vectors for this classification network. The hidden layer has 100 nodes that also use the ReLU activation function. The output layer of the CNN was constructed with a single node, as the decision made by the classifier is binary. The SoftMax function was used at the output node.

Table 4.2 The center frequency and the bandwidth of the gammatone filters

Bandwidth, Hz	**Center Frequency, Hz**
158	50
173	186
276	389
478	690
788	1139
1249	1807
1936	2802

2960	4282

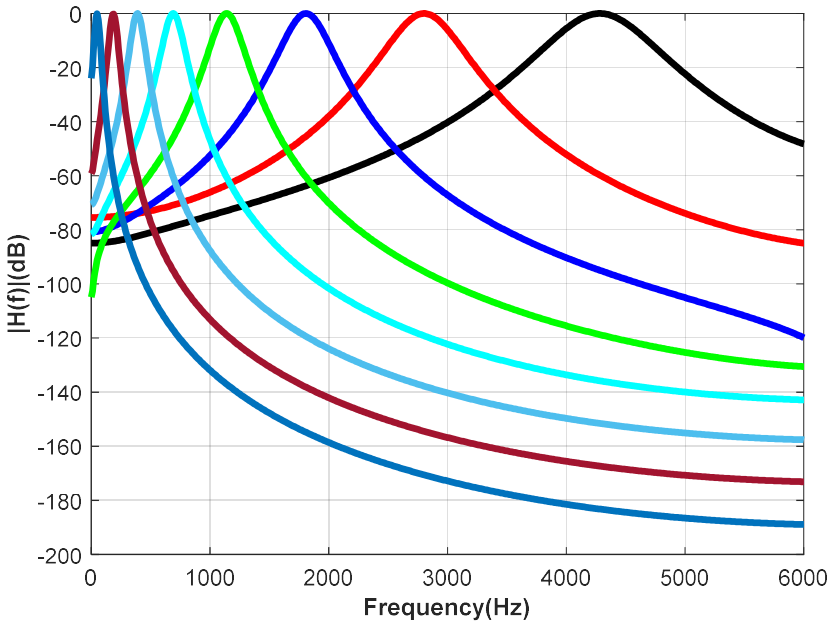


Figure 4.8 The magnitude spectrum of the gammatone filter bank.

4.3 Results

To evaluate the performance of the proposed system, some statistical parameters have been used [154]-[155]. To investigate the proposed system's performance, 10 simulations were conducted using the first investigated model consisting of bandpass filters. First, the CNN was trained with 100 control and 100 pathological samples. Five-fold cross-validation was used to ensure the accuracy of the training. The simulations were run for enough epochs to achieve a training accuracy of 100%. Once trained, the 100 other control samples and 100 pathological samples were used to test the network's performance. The training, validation, and testing results of the proposed algorithm for the first model are listed in Table 4.3. The table shows that the proposed system's average training, validation, and testing accuracies are 100%, 85.96%, and 77.91%, respectively. The testing performances of the proposed system in terms of TPF, TNF, FPF, and FNF are listed in Table 4.4, and the corresponding classification matrix is shown in Table 4.5. Based on the data presented in Table 4.5, it can be concluded that the proposed system can correctly detect pathological voices, resulting in an accuracy of 76.67% with the first model. On the other hand, the system can detect control (i.e., normal) voices with an accuracy of 79.17%.

Ten more simulations were conducted using the second model consisting of the gammatone filters with the same set of control and pathological samples that were used in the previous simulations. The proposed algorithm's training, validation, and testing results are listed in Table 4.6. This table shows that the average training, validation, and testing accuracies of the proposed system are 100%, 81.98%, and 77.50%, respectively. The testing performances of the proposed method in terms of TPF, TNF, FPF, and FNF are listed in Table 4.7, and the corresponding classification matrix is shown in Table 4.8. Based on the data presented in Tables 4.7- 4.8, it can be concluded that the proposed system can correctly identify pathological voices with an accuracy of 83.30% adopting the second model. On the other hand, the system can detect control (i.e., normal) voices with an accuracy of 71.67%.

Table 4.3 Training and testing accuracies with bandpass filters

Simulation No.	*Accuracy* (%)		
	Training	Validation	Testing
1	100	88.00	79.17
2	100	85.83	79.17
3	100	88.83	75.00
4	100	85.83	79.17
5	100	87.83	83.33
6	100	88.00	79.17
7	100	85.83	75.00
8	100	83.33	75.00
9	100	85.33	79.17
10	100	80.83	75.00
Average	100	85.96	77.91

Table 4.4 Simulation results with bandpass filters

Simulation No.	TPF (%)	TNF (%)	FPF (%)	FNF (%)
1	83.33	75.00	25.00	16.67
2	83.33	75.00	25.00	16.67
3	75.00	75.00	25.00	25.00
4	75.00	83.33	16.67	25.00
5	75.00	91.67	8.33	25.00

6	75.00	83.33	16.67	25.00
7	66.67	83.33	16.67	33.33
8	75.00	75.00	25.00	25.00
9	83.33	75.00	25.00	16.67
10	75.00	75.00	25.00	25.00
Average	76.67	79.17	20.83	23.33

Table 4.5 The classification matrix for the bandpass filter model

	Prediction (%)	
Actual	Control	Pathology
Control	79.17 (TNF)	20.83 (FPF)
Pathology	23.33 (FNF)	76.67 (TPF)

The performance comparisons of the two investigated models are listed in Table 4.9. The proposed system performed almost equally in terms of accuracy for both models. The recall was significantly higher for the model with gammatone filters, though precision was greater with bandpass filters. However, the F1 score that considers both recall and precision, was higher for gammatone filters. Also, the NPV was higher for gammatone filters. Hence, it justifies the greater possibility of implementing a signal processing-based pathological voice detection system with gammatone filters, incorporating the functionality of an optimally simulated cochlear implant processing system.

Finally, the performance results of the proposed model were compared with other existing published works, and the comparison is presented in Table 4.10. As listed in this table, the spectrogram audio feature has been used in [96], [126] and the achieved accuracy was 71% for both works. Compared to those works, the proposed system achieved significantly higher accuracies (i.e., 77.9% and 77.5%) for the two studied models. Moreover, the achieved results are challenging as compared with that of [110], where multiple features and mixed pathologies were considered with the children subgroup. Additionally, in [118], a high F-measure (87.75%) was achieved considering vowel samples, but with a speaker-specific identification system.

Table 4.6 Training and testing accuracies with gammatone filters

Simulation No.	*Accuracy* (%)		
	Training	Validation	Testing
1	100	85.00	75.00
2	100	75.83	75.00
3	100	87.83	79.17
4	100	80.83	79.17
5	100	77.83	79.17
6	100	85.00	79.17
7	100	80.83	75.00
8	100	83.33	75.00
9	100	85.00	75.00
10	100	78.83	83.33
Average	100	81.98	77.50

Table 4.7 Simulation results with gammatone filters

Simulation No.	*TPF* (%)	*TNF* (%)	*FPF* (%)	*FNF* (%)
1	83.33	66.67	33.33	16.67
2	83.33	66.67	33.33	16.67
3	83.33	66.67	33.33	16.67
4	83.33	75.00	25.00	16.67
5	83.33	75.00	25.00	16.67
6	83.33	75.00	25.00	16.67
7	83.00	75.00	25.00	16.67
8	75.00	75.00	25.00	25.00
9	91.67	75.00	25.00	8.33
10	83.33	66.67	33.33	16.67
Average	83.30	71.67	28.33	16.67

Table 4.8 The classification matrix for the gammatone filter model

	Prediction (%)	
Actual	Control	Pathology
Control	71.67 (TNF)	28.33 (FPF)
Pathology	16.67 (FNF)	83.30 (TPF)

Table 4.9 Performance comparison of two system models with gammatone and bandpass filters

	System Model	
Measures	Bandpass Filters	Gammatone Filters
Accuracy	77.90	77.50
Precision	78.60	74.60
Recall/Sensitivity	76.70	83.30
F1 score	77.60	78.70
NPV	77.20	81.10

Table 4.10 The performance comparisons with some published voice pathology detection systems

Research Works	Phonemes	Pathological Condition	Features	Tools	Accuracy/ F1 score
Tae Jun [96]	General voice samples	Neoplasm, phono-trauma, vocal palsy	Mel-spectrogram	DNRNN	71%
V. Sellam [110]	Tamil phrases	Multiple voice disorders	Signal energy, pitch, formant frequencies, mean square residual signal, reflection coefficients, jitter and shimmer	SVM RBFNN	Accuracy: 91% (RBFNN) 83% (SVM)
A. Sassou [118],	Japanese Vowel	Roughness, breathiness, asthenia, and strain	HLAC	FFNN, AR-HMM	F-measure: 87.25% for speaker-based identification.
H. Wu [126]	Vowels	Reinke's edema, laryngitis, leukoplakia, recurrent laryngeal, nerve paralysis, vocal fold carcinoma, vocal fold paralysis	Spectrogram	CNN, CDBN	Accuracy: 71%, 77%

Proposed Method	Speech	Laryngitis	Cochlear Simulation Model-1, Cochlear Simulation Model-2	Cochlear implant processing system and CNN	F1 score: 77.6%, 78.7% Accuracy: 77.9%, 77.5%

4.4 Conclusion

This Chapter presented a novel, non-invasive pathological voice detection system considering a cochlear simulation model. Two models have been considered in this work. One model uses a bank of bandpass filters, and the other uses gammatone filters. It has been shown that the gammatone filter is more suitable for voice pathology identification through the signal processing steps involved in the cochlear implants. It has also been demonstrated that the gammatone filters with $f_0/b = 9$ are the optimum choice for this purpose. The speech samples have been processed using these two models and the processed signals were applied to the input of a CNN, which acted as a binary classifier to detect pathological voices. It is a challenging issue to consider suitable features extracted from the speech samples. In general, no single feature or feature vector is well-accepted to provide the best accuracy. This novel technique eliminates acoustic feature extraction from the speech samples before applying the classification algorithm. The simulation results presented in this work have shown that the proposed system achieved almost equal accuracy by using the two proposed models. However, the higher F1 score for the model with gammatone filters illustrates its better applicability for pathological voice identification through the cochlear implant simulation model.

CHAPTER 5
A STUDY OF USING COUGH SOUNDS AND DEEP NEURAL NETWORK FOR THE EARLY DETECTION OF COVID-19

The current clinical diagnosis of COVID-19 requires person-to-person contact, needs variable time to produce results, and is expensive. It is even inaccessible to the general population in some developing countries due to insufficient healthcare facilities. Hence, a low-cost, quick, and easily accessible solution for COVID-19 diagnosis is vital. This Chapter presents a study that involves developing an algorithm for automated and noninvasive diagnosis of COVID-19 using cough sound samples and a deep neural network. The cough sounds provide essential information about the behavior of glottis under different respiratory pathological conditions. Hence, the characteristics of cough sounds can identify respiratory diseases like COVID-19. The proposed algorithm consists of three main steps (a) extraction of acoustic features from the cough sound samples, (b) formation of a feature vector, and (c) classification of the cough sound samples using a deep neural network. The output from the proposed system provides a COVID-19 likelihood diagnosis. In this work, three acoustic feature vectors have been considered, namely (a) time-domain, (b) frequency-domain, and (c) mixed-domain (i.e., a combination of features in both time-domain and frequency-domain). The performance of the proposed algorithm is evaluated using cough sound samples collected from healthy and COVID-19 patients. The results show that the proposed algorithm automatically detects COVID-19 cough sound samples with an overall accuracy of 89.2%, 97.5%, and 93.8% using time-domain, frequency-domain, and mixed-domain feature vectors, respectively. The proposed algorithm, coupled with its high accuracy, demonstrates that it can be used for quick identification or early screening of COVID-19. The obtained results have been compared with that of some state-of-the-art works.

5.1 Introduction

According to the global database maintained by John Hopkins University, more than 270 million COVID-19 (and its variants) cases and 5.3 million deaths have been reported till December 13, 2021 [177]. Social distancing, wearing masks, widespread testing, contact tracing, and massive vaccination are all recommended by the World Health Organization (WHO) to reduce the spreading of this virus [178]. To date, reverse transcription-polymerase chain reaction (RT-PCR) is considered the gold standard for

testing coronavirus [179]. However, the RT-PCR test requires person-to-person contact to administer, needs variable time to produce results, and is still unaffordable to most global populations. Sometimes, it is unpleasant to the children. Not least, this test is not yet accessible to the people living in remote areas, where medical facilities are scarce [180]. Alarmingly, the physicians suspect that the general people refuse the COVID-19 test in fear of stigma [181].

Governments worldwide have initiated a free massive testing campaign to stop the spreading of this virus, and this campaign is costing them billions of dollars per day at the average rate of $23 per test [182]. Hence, easily accessible, quick, and affordable testing is essential to limit the spreading of the virus. The COVID-19 detection method, using human audio signals, can play an important role here.

Researchers and clinicians have suggested using the recordings of speech, breathing, and cough sounds to detect various diseases. As mentioned in Chapter 2 that the speech samples can help clinicians to detect several diseases, including asthma, Alzheimer's disease, Parkinson's disease, depression, schizophrenia, autism, head or neck cancer, and emotional expressiveness of breast cancer patients. Among these diseases, respiratory diseases like asthma have some similarities to COVID-19. An extensive investigation on the detection of asthma using audio signal processing can be found in [24], [183]-[185]. These works show that asthma causes swollen and inflamed vocal folds, which do not vibrate appropriately during voice generation. Hence, the voice samples of asthma patients differ from that of healthy (i.e., control) patients. For example, it is shown in [24] that asthmatic subjects show longer pauses between speech segments, produce fewer syllables per breath, and spend a more significant percentage of time in voiceless ventilator activity than their healthy counterpart.

Recently, researchers have been suggesting using cough sounds for the early detection of the COVID-19. However, there are still some challenges, as the cough is also a symptom of 30 other diseases [186]-[187]. Hence, it is very challenging to discriminate the cough sound of the COVID-19 patients from that of other patients. In [186], the authors considered three diseases: bronchitis, pertussis, and COVID-19. Investigation of 247 normal cough samples and 296 pathological samples was performed. The authors used a CNN to implement a binary classifier and a multiclass classifier. The binary classifier discriminates pathological cough sounds from normal cough sounds, and the multiclass classifier categorizes the pathologies into one of the

three pathology types. In a similar work [188], the authors considered bronchitis, bronchiolitis, and pertussis. They used a CNN to discriminate against these pathologies.

Various human audio samples, namely, the sustained vowel '/a/', counting (1-20), breathing, and cough samples, have been used in [189]. The authors considered nine acoustic voice features: spectral contrast, MFCCs, spectral roll-off, spectral centroid, mean square energy, polynomial fit, zero-crossing rate, spectral bandwidth, and spectral flatness. They used a random forest (RF) algorithm to discriminate the COVID-19 samples from the control/healthy samples and achieved an accuracy of 66.74%.

In [190] , the authors used large samples (5320 samples) selected from the MIT open voice COVID-19 cough dataset [191]. They extracted the MFCC features from the cough samples and classified them by using a CNN. The network consists of one Poisson biomarker layer and three pre-trained ResNet50s. The results showed that their proposed system achieved an accuracy of 97%.

Cough and breathing sounds have also been used in [192]. In that work, the authors used eleven acoustic features: RMS energy, spectral centroid, roll-off frequencies, zero-crossing rate, MFCC, Δ-MFCC, Δ^2-MFCC, tempo, duration onsets, and period. In addition, they used VGGish (a pre-trained CNN from Google) to classify the samples into COVID-positive/non-COVID, COVID-positive with cough/non-COVID with cough, and COVID-positive with cough/non-COVID asthma with cough. The proposed system achieved an accuracy of 80%, 82%, and 80% for the classification tasks mentioned above.

In [193], the authors used Computational Paralinguistic Challenge (COMPARE) [194] features and extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [195] to discriminate the COVID-19 samples from the healthy samples. These features were extracted by using the OpenSMILE [196] tool kit. The voice samples were collected by using five sentences uttered by the patients. The authors classified the COVID-19 patients into three categories, namely, high, mild, and low. In that study, they used 260 samples, including 52 COVID-19 samples. The authors have used a SVM and achieved an accuracy of 69%.

Three acoustic feature sets have been used in [197]. The first one was the COMPARE acoustic features, which were collected by using the OpenSMILE software. The second one was a combination of acoustic feature sets extracted by freely available software, PRAAT [52] and LIBROSA [198]. The third one was an acoustic

feature set consisting of 1024 embedded features extracted by a deep CNN. The samples used in the investigation comprised of three vowels (i.e., '/a/', '/s/', and '/z/'), cough sounds, six symptomatic questions, and counting from 50 to 80. The authors have used the SVM with the RBF and RF as the classifiers. Experimental results showed an average accuracy of 80% in discriminating the COVID-19 positive patients from the COVID-19 negative patients based on the features extracted from the cough and vowel '/a/' recordings. They also achieved even more accuracy (83%) by evaluating six symptomatic questions.

In [199], the authors used voice features, namely cepstral peak prominence (CPP), HNR, first and second harmonic (H1H2), fundamental frequency and its variations (F0SD), Jitter, Shimmer, and maximum phonation time (MPT) to discriminate the voice samples of the COVID-19 patients from that of the healthy patients. The authors collected the sustained vowel sample '/a/' from 70 healthy and 64 COVID-19 patients of Persian speakers. They revealed significantly higher F0SD, Jitter, shimmer, H1H2, and voice break numbers in the COVID-19 patients than the control/healthy group.

Vowels in '/ah/', snoring consonants in '/z/', cough sound, and counting samples from 50 to 80 have been used in [200]. The authors have used a RNN based expert classifier in work. The authors have used three techniques: pre-training, bootstrapping, and regularization to avoid the over-fitting problem of RNN. They also used the leave-one-speaker-out validation technique to achieve a recall of 78%. In a similar work [201], the authors used the RNN algorithm with long short-term memory (LSTM) to detect the COVID-19 patients. In that investigation, the authors used several features, including spectral centroid, spectral roll-off, zero-crossing-rate, MFCCs, and ΔMFCCs from the cough sound, breathing sound, and voice samples of the COVID-19 patients. The authors used 60 healthy and 20 COVID-19 patients in the work. To improve accuracy, they removed the silence part from the samples using the PRAAT software. As a result, the authors achieved an accuracy of 98.2%, 97.0%, and 77.2% by using breathing, cough, and voice samples, respectively.

In [202], the authors have used the MFCC features of cough, breathing, and voice sounds to discriminate the COVID-19 patients from the non-COVID-19 patients. The authors concluded that the MFCCs of cough and breathing sounds for the COVID-19 patients and non-COVID-19 patients are similar. However, the MFCCs of voice sounds are very distinct between the COVID-19 and non-COVID-19 patients.

A cloud computing and artificial intelligence-based early detection of the COVID-19 patients have been presented in [203]. The authors used three-voice features, namely, HNR, Jitter, and Shimmer. In addition, they used the RBF algorithm as a classifier. The authors suggested that the HNR, Jitter, and Shimmer can be used to differentiate between healthy and asthma patients. They also indicated that the same parameters can be used to discriminate between the healthy and COVID-19 patients.

Recurrence quantification measures in the MFCCs have been introduced in [204] to detect the COVID-19 patients using sustained vowel '/ah/' and cough sounds. The authors have used several classifiers, namely, decision trees, SVM, kNN, RF, and XGBoost. Among these classifiers, they achieved the best results with the XGBoost classifier. That model achieved accuracies of 97% (with an F1 Score of 91%) and 99% (with an F1 Score of 89%) for coughs and sustained vowels, respectively.

In [205], the authors used crowdsourced cough audio samples that were acquired on a smartphone from around the world. They collected three acoustic features: MFCCs, Mel-frequency spectrum, and spectrogram from the cough sounds. The authors used an innovative ensemble classifier model (consisting of three networks) to discriminate the COVID-19 patients from the healthy patients. The highest accuracy achieved was 77.1%.

This work is a preliminary investigation of Artificial Intelligence's (AI's) capability to detect COVID-19 by using acoustic features. The proposed algorithm has been developed based on the available data which is limited. Rigorous testing of the algorithm is required with more data before deploying the algorithm in practice for COVID-19 pre-screening. The main contributions of this work are as follows:

- To develop a novel COVID-19 detection algorithm based on signal processing and a DNN.
- To compute the acoustic features and compare their uniqueness for the cough sound samples of control (i.e., healthy) and COVID-19 patients.
- To form the feature vectors using three domains: time-domain, frequency-domain, and mixed-domain, to investigate the efficacy of these feature vectors.
- To achieve a high classification accuracy (compared to other related works) while avoiding an overwhelming computation burden on the system.

- To use a dropout strategy in the proposed algorithms to make the training process faster and to overcome the overfitting problem.
- To provide a detailed performance analysis of the proposed system in terms of the confusion matrix, Accuracy, Precision, NPV, and F1-Score.

5.2 Background

The human voice generation system mainly consists of lungs, larynx, and articulators. Among them, the lungs are considered the power source of the voice generation system. Respiratory diseases prevent the lungs from working properly and hence, affect the human voice generation system. Respiratory diseases can be classified into two main classes, namely, (a) obstructive and (b) restrictive [206]. Obstructive lung diseases make the pulmonary airways narrow and affect a patient's ability to expel air from the lungs completely. Hence, a significant amount of air remains in the lungs all the time. On the other hand, people suffering from restrictive lung diseases cannot fully expand their lungs to fill them with air. Hence, the lungs fail to fully expand. Some patients may suffer from a combination of both obstructive and restrictive respiratory diseases. Cough is the common symptom of obstructive, restrictive, and combined lung diseases. Hence, cough sounds are considered useful for detecting lung diseases caused by respiratory issues [207].

The COVID-19 is also considered a respiratory disease. Like other respiratory diseases, the COVID-19 can cause the lungs to fill with fluid and get inflamed. As a result, patients can suffer from breathing difficulty and need treatment at the hospital with severe onset. Untreated COVID-19 can progress and lead to acute respiratory distress syndrome (ARDS), a form of lung failure [208]. Although coughing is a common symptom of any respiratory illness, including the COVID-19, recent studies suggest that the COVID-19 cough is characterized by dry, persistent, and hoarse at the earliest stage of coronavirus infected patients. Hence, the cough sound samples of the COVID-19 patients differ from those of other patients suffering from some other respiratory diseases. Human cough samples contain three phases: explosive phase, intermediate phase, and voiced phase [209], as shown in Fig. 5.1. These phases represent the glottal airflow variation in the vocal cord, and they differ depending on the pathological conditions of the patients.

Two segmented cough sound samples are randomly selected from the Virufy database [210] to investigate the differences between the cough sound samples of a

COVID-negative (i.e., healthy/control) subject and a COVID-positive patient. The cough samples of a healthy subject and a COVID-positive subject are shown in Fig. 5.2. This figure demonstrates that the healthy sample is similar to the typical human cough signal presented in Fig. 5.1. However, the cough sound sample of the COVID-19 patient varies significantly from the typical human cough sample. For example, both the intermediate and voiced phases are longer for the COVID-positive patient than for the healthy subject.

Moreover, the signal amplitude during the voiced phase is higher for the COVID-positive patient than for the healthy subject. The amplitudes in the explosive phase also differ between these two cough sound samples, as depicted in Fig. 5.2. The differences mentioned above indicate that the cough sound can be used as a valuable tool to discriminate the COVID-positive patient from the healthy subject. The power spectral densities (PSD) of these two samples are plotted in Fig. 5.3. It is observed in the figure that the healthy cough sound has prominent frequencies of continuous decreasing magnitudes. On the other hand, the COVID-positive cough sound samples do not contain very distinct frequencies.

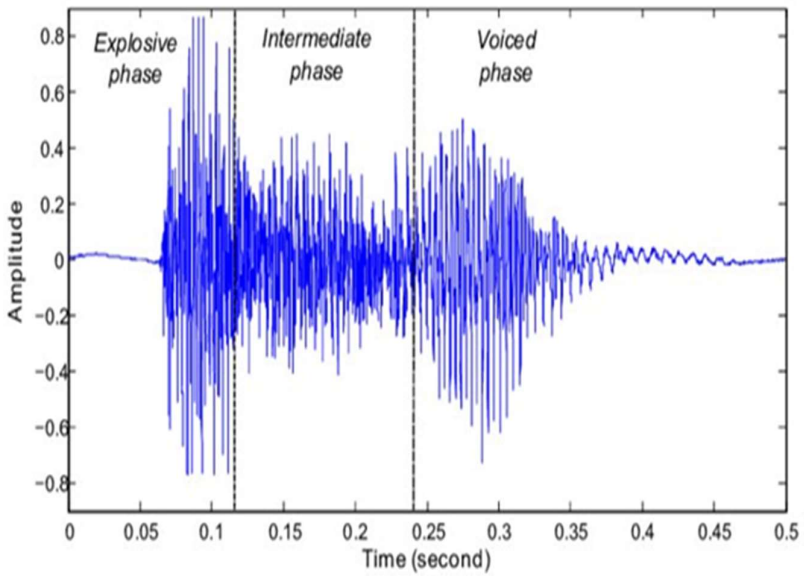


Figure 5.1 A typical cough sound signal phase [209].

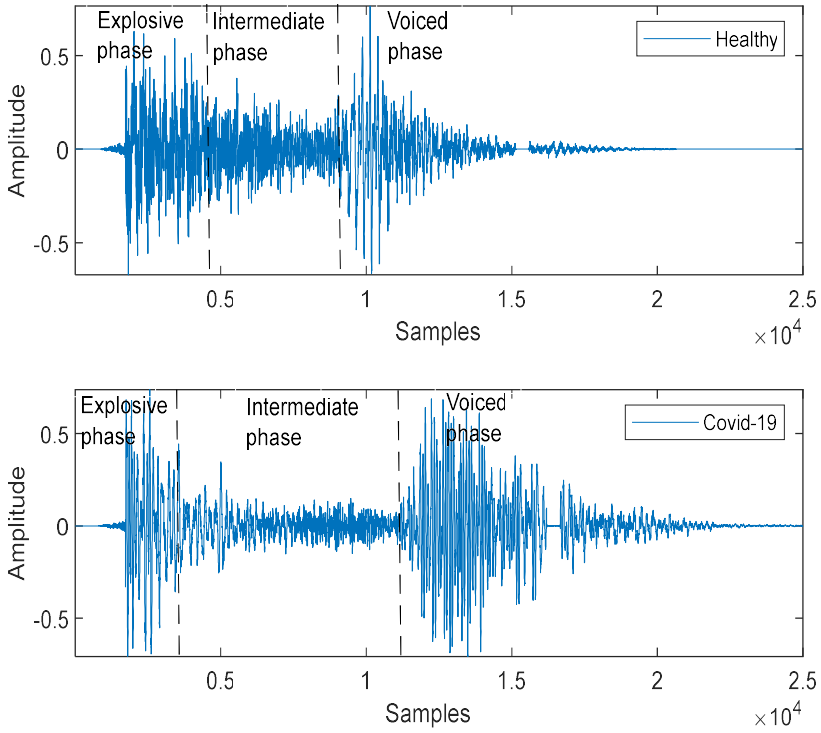


Figure 5.2 Comparison of the cough sounds for a healthy subject and COVID-19 subjects collected from the Virufy database [210].

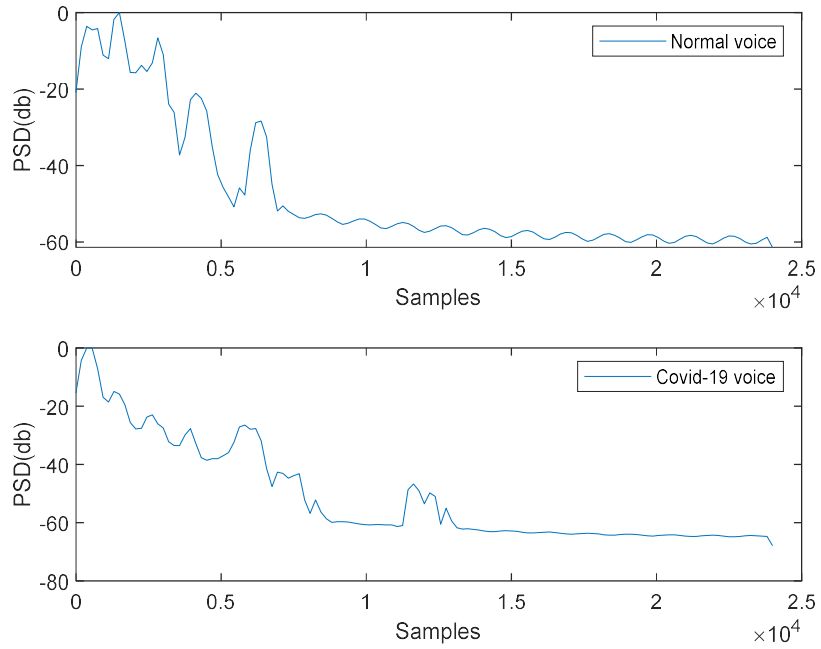


Figure 5.3. Comparison of the power spectral density (PSD) of the cough sounds for a healthy subject and a COVID-19 subject.

5.3 Models, Materials, and Methods

The proposed system model is presented in Fig. 5.4. It consists of four major steps: pre-processing, feature extraction, formation of feature vectors, and classification. The main functions of the pre-processing stage are audio segmentation and windowing. Afterward, the frames are formed. In the next step, the features are extracted from the framed samples. The extracted features are then grouped to form the feature vectors. Finally, the feature vectors are applied as the input to the classifier. The most crucial component of the proposed system is feature extraction (also called the data reduction procedure). It involves extracting features from the cough sound of interest. The main advantage of using features is that the analysis algorithm (i.e., classifier) needs to deal with small and transformed data compared to original voluminous cough sample data.

In practice, acoustic features are extracted, and a feature vector is formed, representing the original data. However, the selection of features and the formation of the appropriate feature vector is an open issue for ongoing research in pattern recognition. In this investigation, 33 acoustic features are considered to form three feature vectors. The acoustic features used in this work can be broadly classified into two major classes: time-domain and frequency-domain features. In this investigation, the cough sound samples are divided into small frames using a rectangular window, and the features are extracted from these frames. These features are explained in the following subsections.

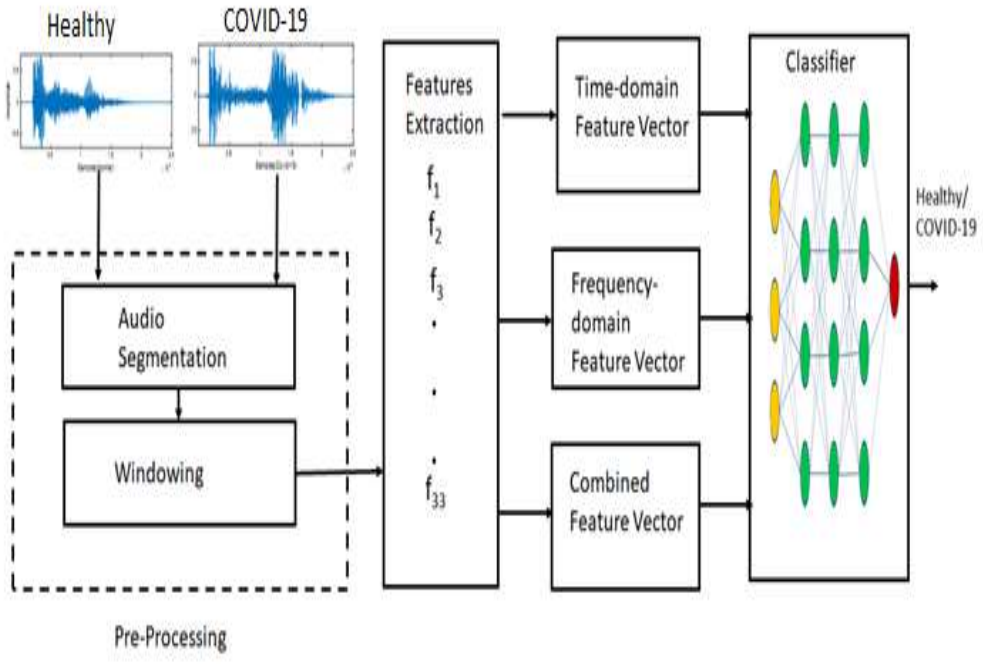


Figure 5. 4 The block diagram of the proposed algorithm.

5.3.1 The Time-domain Features

In this investigation, we consider the following time-domain features: (i) short-term energy, (ii) zero-crossing rate, and (iii) entropy of energy [211]. The short-term energy of the ith frame is calculated by

$$E(i) = \sum_{n=1}^{W_L} |x_i(n)|^2 , \tag{5.1}$$

where, $x_i\ (n)$ is the ith frame, with W_L being the length of the frame. The energy expressed in (5.1) is normalized as

$$E_n(i) = \frac{E(i)}{W_L} = \frac{\sum_{n=1}^{W_L} |x_i(n)|^2}{W_L} , \tag{5.2}$$

The normalized energy content of the COVID-positive and healthy cough sounds is plotted in Fig. 5.5(a). This figure shows that the energy content of both samples are concentrated in a few frames, and they exhibit a high variation over successive frames. However, the energy content of the COVID-positive patient is much higher than that of the healthy subject. It indicates that the cough sample of the COVID-positive patient contains weak phonemes and a short period of silence between two coughs. Hence, the energy content also varies rapidly between two successive frames.

The zero-crossing rate of a cough sound signal can be defined as the rate of sign changes of the movement over the frames. It is calculated by using the following equation

$$Z(i) = \frac{1}{2W_L} \sum_{n=1}^{W_L} |sgn[x_i(n)] - sgn[x_i(n-1)]|, \tag{5.3}$$

where, $sgn(\bullet)$ is the sign function defined by $sgn[x_i(n)] = 1$, when $x_i\ (n) \geq 0$ and $sgn[x_i(n)] = -1$, when $x_i\ (n) < 0$. The zero-crossing rate of the COVID-positive patient and the healthy subject are plotted in Fig. 5.5(b), which shows that the healthy cough sample has a more zero-crossing rate than that of the COVID-positive patient. Since the zero-crossing rate measures the noisiness of a signal, it exhibits a higher zero-crossing rate for the unvoiced part of the cough sound sample and a lower zero-crossing rate for the voiced samples. As shown in Fig. 5.2, the voiced phase of the cough samples

for the COVID-positive patient is longer than that of the healthy subject. Hence, the zero-crossing rate is lower for the COVID-positive patient than for the healthy subject, as depicted in Fig. 5.5(b). The short-term entropy of energy can be interpreted as a measure of the abrupt changes in the energy level of an audio signal. To compute it, we first divide each short-term frame into K sub-frames of fixed duration. Then, for each sub-frame, j, the energy is calculated by using (5.1) and divide it by the total energy, $E_{short\_frame_i}$ of the short-term frame. Then, the sub-frame energy values, e_j, for j =1,2, …, K, is computed as a sequence of probabilities and is defined as

$$e_j = \frac{E_{sub\_frame_j}}{E_{short\_frame_i}}, \tag{5.4}$$

where, $E_{short\_frame_i} = \sum_{k=1}^{K} E_{sub\_frame_k}$. At the final step, the entropy, $H(i)$, is calculated from the sequence e_j by

$$H(i) = -\sum_{j=1}^{K} e_{j.} log_2(e_j), \tag{5.5}$$

The short-term entropy of energy for the COVID-positive patient and the healthy subject are plotted in Fig. 5.5(c). The short-term entropy of energy for the COVID-positive patient is greater than that of the healthy subject for most of the frames. Since the energy content of the COVID-positive patient varies more abruptly than that of the healthy subject, the energy entropy tends to be higher for the COVID-positive patient after frame 20, as shown in Fig. 5.5(c).

5.3.2 Frequency Domain Features

The frequency-domain acoustic features are extracted from the DFT of a signal. The DFT of a frame of audio signal can be expressed as

$$X_i(k) = \sum_{n=1}^{N-1} x_i\,(n) e^{-j\frac{2\pi}{N}nk}, \tag{5.6}$$

where, N is the size of the DFT, $X_i(k)$ is the value of the DFT coefficients, and $k = 1,2,\ldots W_L$. The spectral centroid dictates a noise-robust estimate of the dominant frequency for the cough sound signal that varies over time. It is also called the center

of gravity of the spectrum. The value of the spectral centroid, C_i, of the ith audio frame is calculated by

$$C_i = \frac{\sum_{k=1}^{W_L} kX_i(k)}{\sum_{k=1}^{W_L} X_i(k)} \quad , \tag{5.7}$$

The spectral centroids of the COVID-19 positive and the healthy person are shown in Fig. 5.6(a). It is shown in the figure that the spectral centroids of the cough sound for the healthy person are higher compared to those of the COVID-19 cough samples until approximately frame number 50. The highest value corresponds to the brightest sound practically. Usually, the existence of noise, silence, etc. signifies the lower values of the spectral centroid. This is noticeable for COVID-positive patient as opposed to the healthy person for the range mentioned above. From nearly 50-80 frames, the COVID-positive patient exhibits higher values of the spectral centroid. After frame number 80, both the samples show insignificant spectral components.

The spectral entropy is a measure of irregularities in the frequency domain. The spectral entropy features are computed from the STFT spectrum. Spectral entropy is widely used to detect the voiced regions of an acoustic signal. The flat distribution of silence or noise induces high entropy values. The spectral entropy is computed with the same method that follows to calculate the cough signal's energy entropy. First, the spectrum of the short-term frame is divided into L sub-bands. The energy, E_f, of the fth sub-band, where $f = 0,1,2, \dots , (L-1)$, is normalized by the total spectral energy. The normalized energy is defined as $n_f = \frac{E_f}{\sum_{f=0}^{L-1} E_f}$, $f = 0,1,2, \dots, (L-1)$. Finally, the entropy of the normalized spectral energy, n_f , is computed by

$$H = -\sum_{f=0}^{L-1} n_f . log_2(n_f) \, , \tag{5.8}$$

The spectral entropies of the COVID-positive and the healthy person are shown in Fig. 5.6(b). This figure shows that the spectral entropy of the healthy person is higher than that of the COVID-positive patient for most of the frames. The reason is that the voiced part of the signal contains less spectral entropy than the unvoiced one.

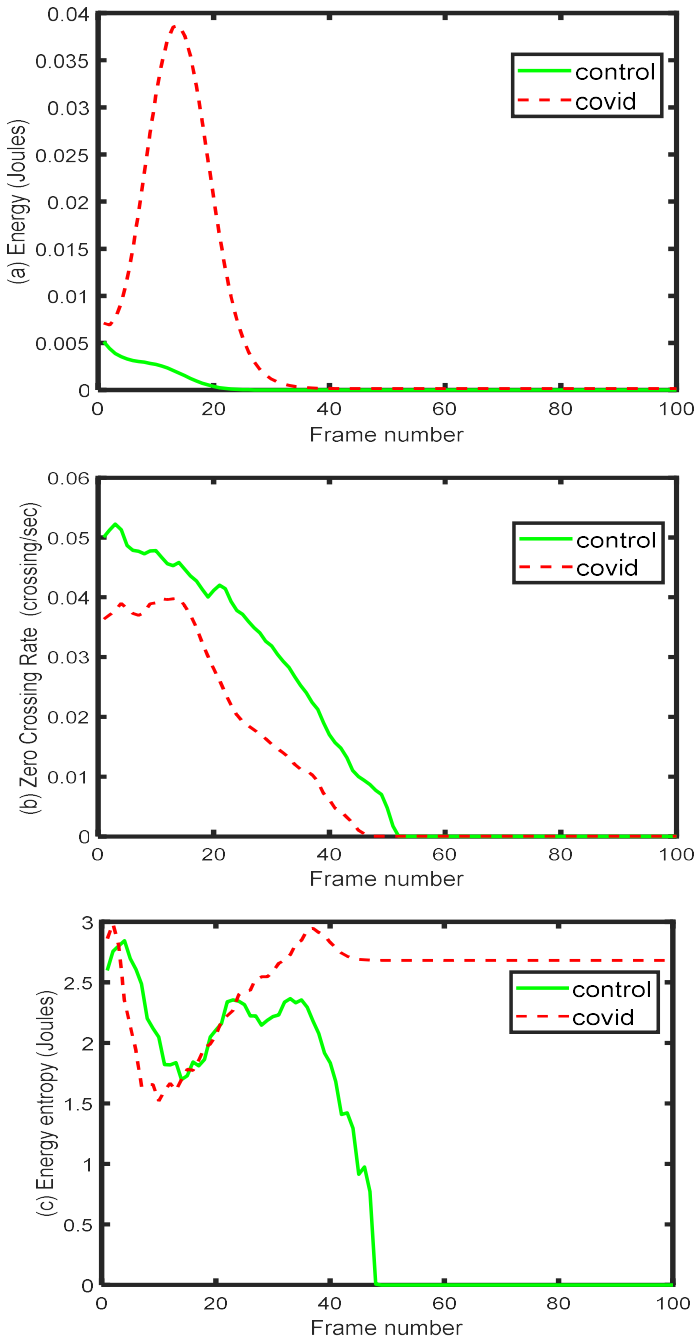


Figure 5.5 The time-domain features (a) short time energy distribution, (b) short time zero-crossing rate, and (c) energy entropy.

The spectral flux measures the spectral change between two successive frames. The spectral flux is computed as the squared difference between the normalized magnitudes of the spectra for the two subsequent short-term windows. It is defined by

$$Fl_{(i,i-1)} = \sum_{k=1}^{W_L}[EN_i\,(k) - EN_{i-1}(k)]^2, \tag{5.9}$$

where, $EN_i(k) = \frac{X_i(k)}{\sum_{l=1}^{W_L} X_{i(l)}}$. The spectral fluxes of the cough sample for the COVID-positive and the healthy person are plotted in Fig. 5.6(c). The magnitude of the spectral flux is higher for the healthy person compared to the COVID-positive patient for most frames. The reason is the more frequent local spectral changes in the healthy cough samples than in the COVID-positive ones. This indicates more rapid spectral alternation among phonemes in the healthy cough sample than in the COVID-positive patient.

The spectral roll-off is the frequency below which a certain percentage (usually around 90%) of the magnitude distribution of the spectrum is concentrated. Therefore, if the mth DFT coefficient corresponds to the spectral roll-off of the ith frame, then it satisfies the following equation

$$\sum_{k=1}^{m} X_i(k) = C \sum_{k=1}^{W_L} X_i(k)\,, \tag{5.10}$$

where, C is the adopted percentage (user parameter). The spectral roll-off frequency is usually normalized by dividing it with W_L, so that it takes values between 0 and 1. The spectral roll-offs of the cough samples for the healthy person and the COVID-positive patient are shown in Fig. 5.7(a). It can be easily observed that the cough samples of the healthy person show a higher spectral roll-off value than that of the COVID-positive patient for most of the frames. It means that the cough sample of the healthy person has a wider spectrum compared to that of the COVID-positive patient.

The MFCCs are also used to form the feature vector. In this work, 13 MFCCs have been considered. The plots for the arbitrarily chosen 7th coefficient of the MFCCs for both the healthy cough samples and COVID-positive cough samples are shown in Fig. 5.7(b). It is shown in the figure that the magnitude of the 7th MFCC coefficient is higher for the COVID-positive cough sample compared to that of the healthy cough sound for most of the frames.

The chroma vector used in this work is a 12-element representation of spectral energy. The chroma vector is computed by grouping the DFT coefficients of a short-term window into 12 bins. Each bin represents the 12 equal-tempered pitch classes of semitone spacing. Also, each bin produces the mean of the log-magnitudes of the respective DFT coefficients defined by

$$v_k = \sum_{n \in S_k} \frac{X_i(n)}{N_k},\ k \in 0,\ldots,\ 11, \tag{5.11}$$

where, S_k is a subset of the frequencies that correspond to the DFT coefficients and N_k is the cardinality of S_k. In the context of a short-term feature extraction procedure, the chroma vector v_k is usually computed on a short frame basis. This results in a matrix V, with elements $V_{k,i}$, where indices k and i represent pitch-class and frame-number, respectively. The chroma vector plots of the healthy and the COVID-positive cough samples are shown in Fig. 5.7(c). It is shown that the chroma vector of the healthy person shows one dominant coefficient, and the rest of the coefficients are of small magnitudes. On the other hand, the chroma vector of the COVID-positive cough sample is nosier and does not have any dominant coefficient. In addition, the chroma vector of the cough sample for the COVID-positive patient does not contain any nonzero coefficient.

The autocorrelation function for the ith frame is computed by

$$R_i(m) = \sum_{n=1}^{W_L} x_i(n) x_i(n-m)\ , \tag{5.12}$$

Actually, $R_i(m)$ is the correlation of the ith frame with itself at time lag, m. Then the autocorrelation function is normalized as

$$\mathrm{r}_i(m) = \frac{R_i(m)}{\sqrt{\sum_{n=1}^{W_L} x_i(n)^2 \sum_{n=1}^{W_L} x_i(n-m)^2}}, \tag{5.13}$$

Afterward, the maximum value of r_i, i.e., the harmonic ratio is calculated as

$$HR_i = max\{\mathrm{r}_i(m)\}, \tag{5.14}$$

where , $T_{min} \leq m \leq T_{max}$, T_{min} and T_{max} are the minimum and maximum allowable values of the fundamental period. Here, T_{max} is often defined by the user, whereas T_{min}

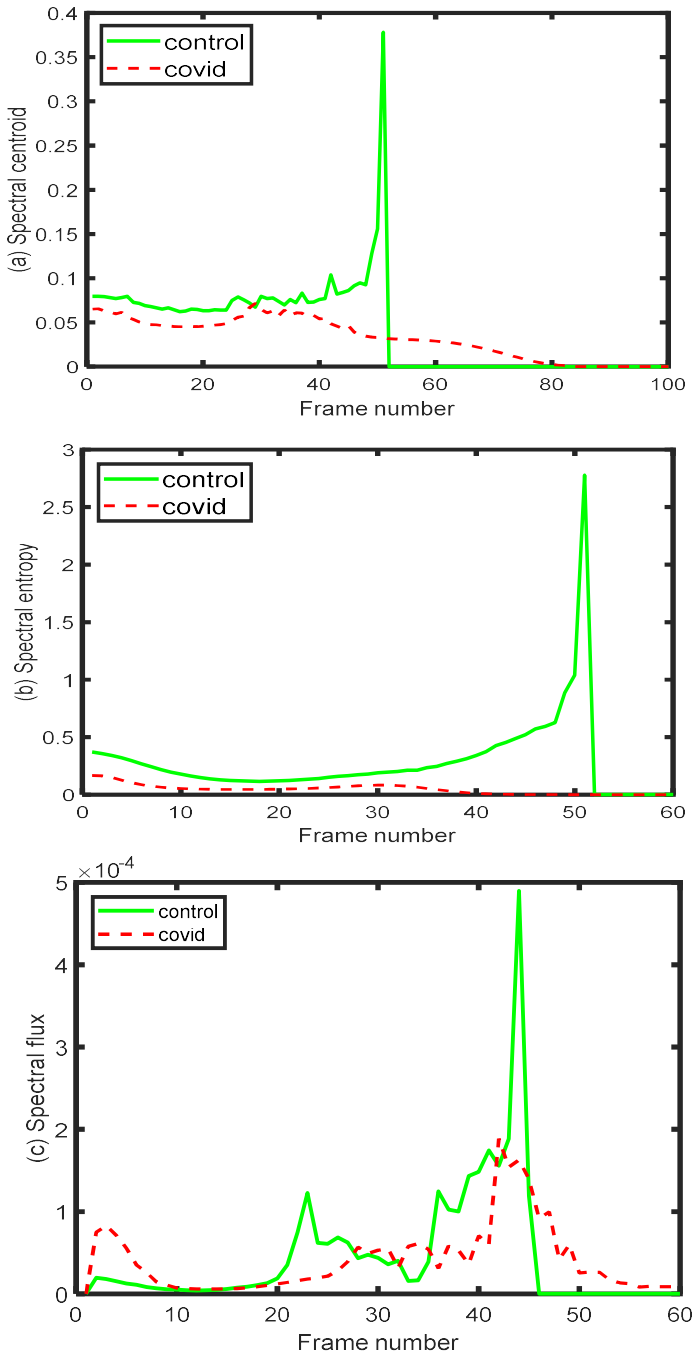


Figure 5.6 The frequency-domain features (a) spectral centroid, (b) spectral entropy, and (c) spectral flux.

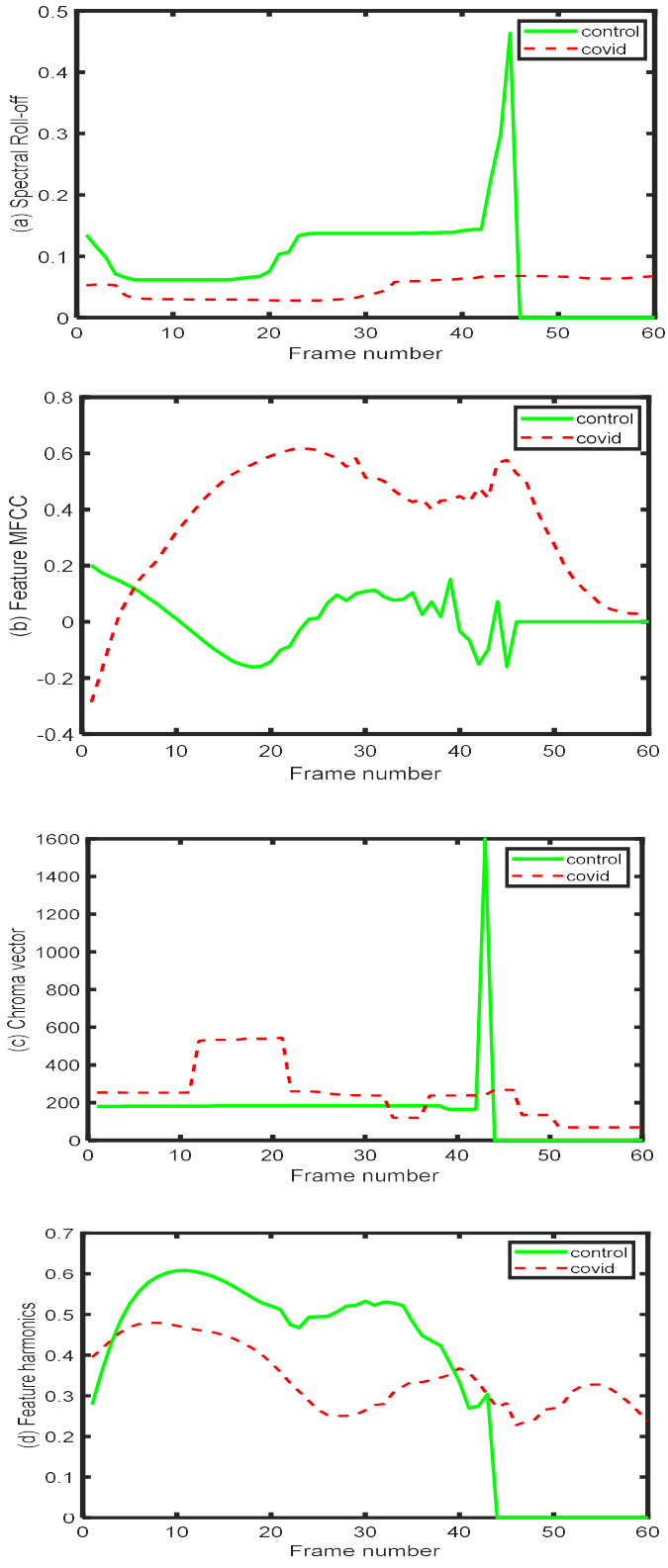


Figure 5.7 The frequency-domain features (a) spectral roll-off, (b) MFCC coefficient, (c) chroma vector, and (d) feature harmonics.

usually corresponds to the lag in time for which the first zero crossing of the $r_i(m)$ occurs. The plots for the harmonic ratio of the healthy and the COVID-positive patients are shown in Fig. 5.7(d). It is depicted in the figure that the harmonic ratio of the cough sample for the healthy person is higher for most of the frames. However, the harmonic ratio shows nonzero values for all analysis frames of the COVID-positive cough samples. On the other hand, the harmonic ratio of the healthy person has zero values for some of the analysis frames.

In this work, the cough samples collected from the Virufy database [210] are used. The Virufy is a volunteer-run organization, which has built a global database to identify the COVID-19 patients using AI. The database contains both clinical data and crowdsourced data. The clinical data is accurate because it was collected and reprocessed at a hospital following a standard operating procedure (SOP). Qualified physicians monitored the whole process of data collection. The patients were confirmed as healthy persons (i.e., COVID-19 negative) and COVID-19 patients (i.e., COVID-positive) by using the PCR test, and the data was labeled accordingly. The database also contains the patients' information, including age, gender, and medical history. Virufy provided 121 segmented cough samples from these 16 patients. The Virufy database contains both the original cough audio recordings and the segmented version of the cough sounds. The segmented coughs were created by identifying the periods of relative silence in the recordings and separating cough samples based on those silences. The segments with no coughing or having too much background noise were removed. The crowdsourced data, maintained by Virufy, is diverse and donated by patients from multiple countries. This database is significantly increasing in volume over time as more people are contributing their cough samples. In this work, only the clinically collected cough samples are used as they are more authentic than crowdsourced data and, also, hence the segmented cough samples have been used in this work.

A DNN discriminates the COVID-19 cough sound samples from the healthy cough sound samples, as shown in Fig. 5.4. The DNN model presented in [152] is used and modified to implement our system. The DNN used in the network consists of three hidden layers. Each hidden layer consists of 20 nodes.

The network has 500 input nodes for the matrix input. It has only one output node as the decision is binary. The output node employs the SoftMax activation function, whereas the hidden nodes consist of the sigmoid function. One of the limitations of the DNN is that they are vulnerable to overfitting. This problem worsens as the network

includes more nodes. To solve the overfitting problem, a dropout algorithm is employed. This algorithm trains only some of the randomly selected nodes rather than all the entire network nodes. The dropout effectively prevents overfitting as it continuously alters the nodes and weights in the training process. In this work, a dropout ratio of 10% and 20% are used for the input and hidden layers, respectively.

5.4 Simulation Results and Discussion

The samples used in the work are distributed into three parts: 70% are for training the DNN, the remaining 30% into validation, and testing with a ratio of 2:1. Five-fold validation is used. The data samples and patient information are listed in Table 5.1. The proposed system's training, validation, and testing results with the three feature vectors are listed in Table 5.2.

First, the time-domain feature vector is used that has three acoustic features, namely, zero-crossing rate, energy, and energy entropy. Then, the DNN (with five-fold cross-validation) is trained, and the system is tested with the time-domain feature vector. The results are shown in Table 5.2, with an average training accuracy of 100%, validation accuracy of 93.27%, and testing accuracy of 89.20%. The confusion matrix of the time-domain feature vector is provided in Table 5.3. Based on the data presented in Table 5.3, it can be concluded that the DNN can correctly detect the COVID-positive cough sound samples with an accuracy of 86.67% by using the time-domain features. On the other hand, it can detect healthy cough samples with an accuracy of 91.67%.

Simulations are repeated by using the frequency-domain feature vector. As mentioned before, the features considered are spectral centroid, spectral entropy, spectral flux, spectral roll-offs, MFCCs, and chroma vector. The training, validation, and testing results are also listed in Table 5.2. The data shows that the DNN achieves training accuracy of 100%, validation accuracy of 98.50%, and testing accuracy of 97.50% by using the frequency-domain feature vector. It can be concluded that the testing accuracy of the frequency-domain feature vector is higher than that of the time-domain feature vector. The confusion matrix of the frequency-domain feature vector is presented in Table 5.4, which shows that the frequency-domain feature vector boosts the DNN’s ability to detect the COVID-positive cough sound samples with an accuracy of 95%. Moreover, the DNN can detect healthy samples with an accuracy of 100%. Both parameters are higher than those of the time-domain feature vector.

Table 5.1 Data samples

Sample	Corona test	Age	Gender	Medical history	Reported symptoms	Cough file name
1	Negative	53	Male	None	None	neg-0421-083-cough-m-53.mp7
2	Positive	50	Male	Congestive heart failure	Shortness of breath	pos-0421-084-cough-m-50.mp3
3	Negative	43	Male	None	Sore throat	neg-0421-085-cough-m-43.mp3
4	Positive	65	Male	Asthma/chronic lung disease	Shortness of breath, new or worsening cough	pos-0421-086-cough-m-65.mp3
5	Positive	40	Female	None	Sore throat, loss of taste, loss of smell	pos-0421-087-cough-f-40.mp3
6	Negative	66	Female	Diabetes with complication	None	neg-0421-088-cough-f-66.mp3
7	Negative	20	Female	None	None	neg-0421-089-cough-f-20.mp3
8	Negative	17	Female	None	Shortness of breath, sore throat, body aches	neg-0421-090-cough-f-17.mp3
9	Negative	47	Male	None	New or worsening cough	neg-0421-091-cough-m-47.mp3
10	Positive	53	Male	None	Fever, chills, or sweating, shortness of breath, new or worsening cough, sore throat, loss of taste, loss of smell	pos-0421-092-cough-m-53.mp3
11	Positive	24	Female	None	None	pos-0421-093-cough-f-24.mp3
12	Positive	51	Male	Diabetes with complication	Fever, chills, or sweating, new or worsening cough, sore throat	pos-0421-094-cough-m-51.mp3
13	Negative	53	Male	None	None	neg-0422-095-cough-m-53.mp3
14	Positive	31	Male	None	Shortness of breath, new or worsening cough	pos-0422-096-cough-m-31.mp3
15	Negative	37	Male	None	None	neg-0422-097-cough-m-37.mp3
16	Negative	24	Female	None	New or worsening cough	neg-0422-098-cough-f-24.mp3

Table 5.2 Training and testing accuracy of the feature vectors

Feature Vector	Training Accuracy (%)	Validation Accuracy (%)	Testing Accuracy (%)
Time-domain feature vector	100	93.27	89.20
Frequency-domain feature vector	100	98.50	97.50
Mixed feature vector	100	96.37	93.80

Table 5.3 The confusion matrix of the time-domain feature vector

	Prediction (%)	
Actual	Healthy	COVID-19
Healthy	91.67 (S^-)	8.33 (FPF)
COVID-19	13.33 (FNF)	86.67 (S^+)

Table 5.4 The confusion matrix of the frequency-domain feature vector

	Prediction (%)	
Actual	Healthy	COVID-19
Healthy	100.00 (S^-)	0.00 (FPF)
COVID-19	5.00 (FNF)	95.00 (S^+)

Lastly, time-domain and frequency-domain features are combined to form a mixed-feature vector. The training, validation, and testing accuracies for the mixed feature vector are listed in Table 5.2. The achieved training, testing, and validation accuracies are 100%, 96.37%, and 93.80%, respectively. The confusion matrix of the mixed-feature vector is presented in Table 5.5. The DNN can detect COVID-positive cough sound samples with an accuracy of 93.34%. On the other hand, it can detect the healthy cough sound samples with an accuracy of 94.17%.

Table 5.5 The confusion matrix of the mixed feature vector

	Prediction (%)	
Actual	Healthy	COVID-19
Healthy	94.17 (S^-)	5.82 (FPF)
COVID-19	6.67 (FNF)	93.34 (S^+)

The performances of the proposed system in terms of Accuracy, Precision, F1 Score, and NPV for the time-domain feature vector, frequency-domain feature vector, and mixed domain feature vector are listed in Table 5.6. This table shows that the proposed system achieves the highest accuracy of 97.5% using the frequency-domain feature vector. On the other hand, the lowest accuracy of 89.2% is achieved using the time-domain feature vector. The other performance scores, including precision, F1 Score, and NPV, are the highest in the frequency-domain feature vector.

Table 5.6 The performance comparison

Measures	Time-domain feature vector	Frequency- domain feature vector	Mixed feature vector
Accuracy	0.892	**0.975**	0.938
Precision/ PPV	0.912	**1.000**	0.941
F1 Score	0.889	**0.974**	0.937
NPV	0.873	**0.952**	0.934

Cough is regarded as a natural defense mechanism of some respiratory disorders, including COVID-19. The human audible hearing range impaired existing subjective clinical approaches of cough sound analysis [212]. Exploration of noninvasive diagnostic approaches well above the audible frequency range (i.e., 48000 Hz) used for sample data can overcome this limitation demonstrated in this study. The non-stationary characteristics of cough sound impose additional challenges for signal processing-based approaches. Also, cough patterns show variability in human subjects under the same pathological state. The cough features that are closely tied to the intensity levels as in the time domain can have dissimilarity for the identical pathology. The cough sound is characterized by the fundamental frequency and significant harmonics when pathology is involved. The restriction of airways causes turbulence in the cough sound that constitutes the harmonics [209]. More realistically, a method that captures both time and frequency changes over the cough samples should associate the investigated respiratory disorder, i.e., COVID-19, with greater accuracy. The best diagnostic performance of the frequency-domain feature vector

in Table 5.6 justifies that the cough features distributed in the frequency domain should possess greater significance.

Finally, the performance of the proposed system is compared with other related works available in the literature, as listed in Table 5.7. The comparison table shows that the proposed system achieves a higher accuracy of 97.5% with the frequency-domain feature vector using the cough sound samples compared to [201]. The system achieves even higher accuracy with the time-domain and mixed-feature vector than the works published in [189], [192]-[193], and [213].

5.5 Research Applicability

Since the publicly available databases are restricted to COVID-positive and COVID-negative (i.e., healthy/control) cases, this study focuses on discriminating COVID-19 cough sound from the healthy cough sound. However, the proposed algorithm can have the possibility to differentiate pathological cough sounds into distinct pulmonary/respiratory diseases, including COVID-19, asthma, bronchiectasis, etc. The pathophysiology and acoustic property of cough sounds can provide significant information in the frequency domain to characterize them for multi-classification purposes. Asthma causes the airways of the patient to be inflamed and narrower. On the other hand, bronchiectasis damages the airways and widens them abnormally. Few randomly selected cough sound samples of some respiratory disorders are investigated in [214]. The samples available in [215] are not sufficient to apply the proposed deep learning-based algorithm. One sample of asthma and bronchiectasis cough sound each is shown in Fig. 5.8 to demonstrate their uniqueness in the time domain. Bronchiectasis cough sound has longer cough sequences compared to asthma cough sound. Additionally, the bronchiectasis cough sound demonstrates more flow spikes than the asthmatic cough sound [209]. These flow spikes indicate more severe inflammation in bronchiectasis patients than in an asthmatic patient. When comparing Fig. 5.2 and Fig. 5.8, it can be concluded that the explosive, intermediate, and voiced phases are very distinct in the COVID-19 cough sample; however, these phases are hardly visible in asthma and bronchiectasis cough sounds.

Table 5.7 The performance comparison with existing works

Research Work	Samples	Phonemes	Features	Classifier	Accuracy
N. Sharma [189]	Healthy and COVID-positive: 941	Cough, Breathing, Vowel, and Counting (1:20)	Spectral contrast, MFCC, Spectral roll-off, Spectral centroid, Mean square energy, Polynomial fit, zero-crossing rate, Spectral bandwidth, and Spectral flatness	Random Forest	66.74%
C. Brown et al. [192].	COVID-positive: 141 Non-COVID: 298 COVID-positive with Cough: 54 Non-COVID with cough: 32 Non-COVID asthma: 20	Cough, and Breathing	RMS energy, Spectral centroid, Roll-off frequencies, Zero-crossing, MFCC, Δ-MFCC,Δ^2 MFCC, Duration Tempo Onsets, Period	CNN	80%
J. Han [193]	COVID-Positive: 52 Healthy: 208	Voice	Computational Paralinguistic Challenge (COMPARE) feature set, and Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)	SVM	69%
A.Hassan [201]	COVID-Positive: 20 Healthy: 60	Breathing, Cough, Voice	Spectral centroid, Roll-off frequencies, Zero-crossing, MFCC, and Δ-MFCC	RNN	98.2% (Breathing) 97% (Cough) 88.2%(Voice)
V. Espotovic [213]	COVID-Positive: 84 COVID-Negative: 1019	Voice, Cough Breathing	Wavelet	Ensemble Boosted	88.52%
Proposed System (time-domain)	COVID-Positve:50 Healthy: 50	Cough	Zero-crossing rate, energy, and energy entropy	DNN	89.2%
Proposed System (Frequency-domain)	COVID-Positve:50 Healthy: 50	Cough	Spectral centroid, spectral entropy, spectral flux, spectral roll-offs, MFCC, and chroma vector	DNN	97.5%
Proposed System (Mixed- feature)	COVID-Positve:50 Healthy: 50	Cough	zero-crossing rate, energy, energy entropy, spectral centroid, spectral entropy, spectral flux, spectral roll-offs, MFCC, and chroma vector	DNN	93.8%

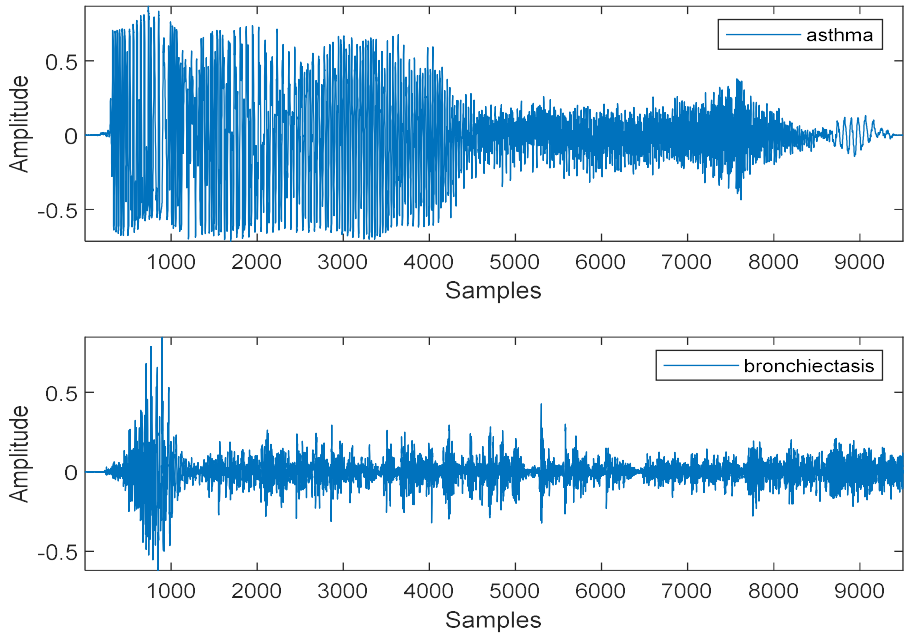


Figure 5.8 The cough sound samples of asthma and bronchiectasis.

As demonstrated in this study, some of the frequency-domain features of COVID-19, asthma, and bronchiectasis cough samples are plotted in Fig. 5.9 to show their uniqueness. The spectral entropy of the bronchiectasis sample is much higher for most of the frame compared to COVID-19 and asthma cough samples. The other features including spectral flux, MFCC, and feature harmonics are also non-identical for the mentioned three respiratory disorders. The distinct differences for the frequency domain features indicate that the proposed algorithm can also be applied to differentiate COVID-19 from asthma and bronchiectasis cough samples, provided a good number of datasets are available for each class.

5.6 Conclusion

In this Chapter, a DNN-based study for the early detection of COVID-19 patients has been presented using cough sound samples. The study proposed a system that extracts the acoustic features from the cough sound samples and forms three feature vectors.

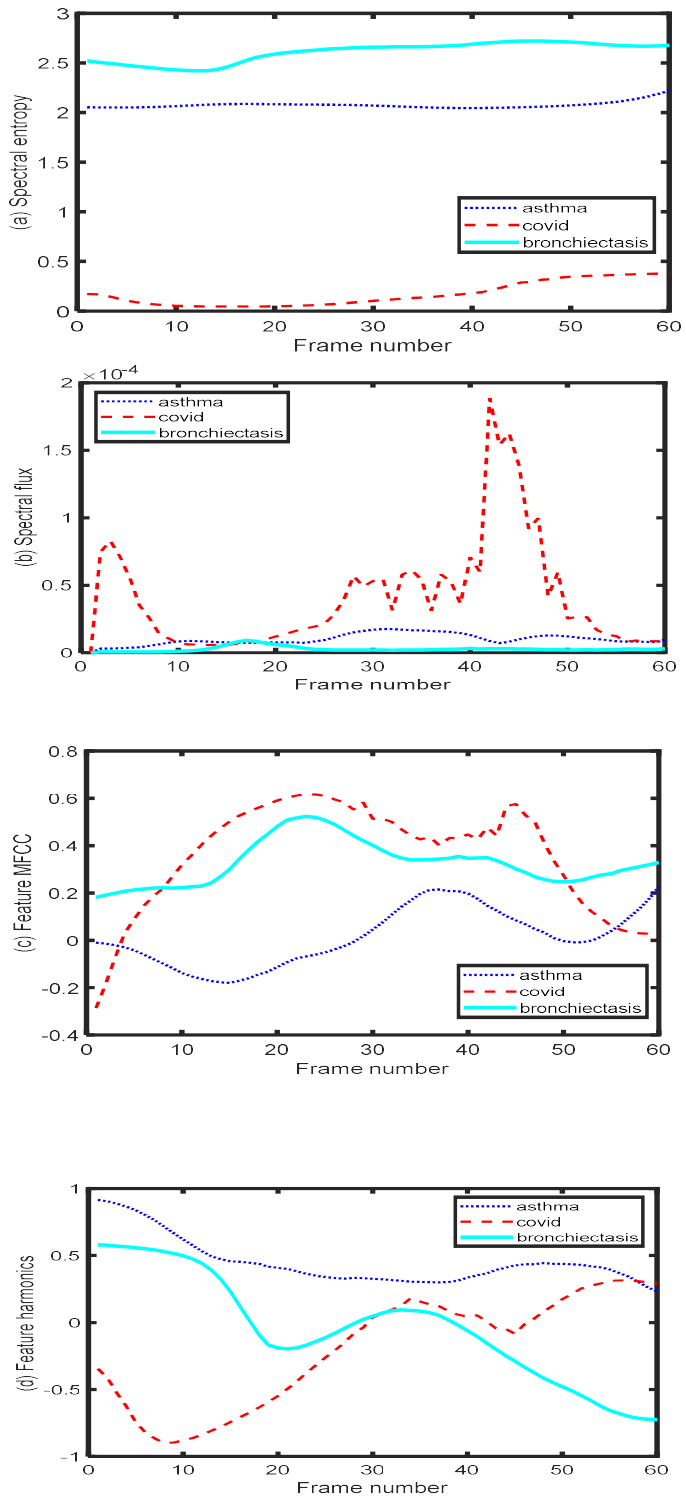


Figure 5.9. The frequency domain features of (a) Spectral entropy, (b) Spectral flux, (c) MFCC coefficient (6^{th}), and (d) Feature harmonics for COVID-19, asthma, and bronchiectasis cough samples.

A rigorous, in-depth investigation has been provided in this work to show that the cough sound samples can be a valuable tool to discriminate the COVID-19 patient from other healthy cough samples for preliminary assessment instead of using the RT-PCR test. In this work, it has been shown that some acoustic features are unique in the cough sound samples of the COVID-19 patients and hence can be used by a classifier like DNN to discriminate them from the healthy cough sound samples successfully. However, there has always been an argument about selecting the appropriate acoustic features for the classifications. The major challenges are (a) to decide whether to use a single feature (like MFCC, spectrogram, etc.) or feature vector, (b) to select the appropriate combination of acoustic features to form the feature vector, and (c) to choose the appropriate domain (i.e., time-domain, frequency-domain, or both). Three feature vectors have been investigated in this work to address this issue. It was shown and justified that the frequency-domain feature vector has provided the highest accuracy compared to the time-domain or mixed-domain feature vector.

The performance of the proposed system has been compared with those of other existing state-of-the-art methods that are presented in the literature for the diagnosis of COVID-19. This accessible and noninvasive pre-diagnosis technique can enhance the screening of all COVID-positive cases, including asymptomatic and pre-symptomatic cases. Also, early diagnosis can help them to stay in touch with healthcare providers for a better prognosis to avoid the critical consequences of COVID-19.

CHAPTER 6
COCHLEAGRAM TO DETECT VOICE PATHOLOGY: AN AUDITORY PERCEPTUAL ANALYSIS

This Chapter presents a deep learning-based, non-invasive technique to manifest and discriminate dysphonic voices through auditory perceptual analysis. The spectral images provide the dynamic characteristics of the voice signal in time and frequency domains. However, extracting the predominant spectral features from the voice samples by transforming them into suitable image forms is still challenging. This Chapter suggests employing cochleagram image to unveil the detailed spectral content of the voice samples and hence it can be considered a valuable tool to recognize dysphonic voice. Both sustained vowel (‘/a/’) and sentence voice samples are considered to include phonation, respiration, and resonance of the vocal characteristics. Treating male and female voice samples separately is also suggested to eliminate gender bias, as they have structurally different vocal tracts, pharynx, and oral cavities. Finally, a pre-trained CNN is used as a classifier to overcome the limitation of sparse data samples. The simulation results show that the cochleagram, coined with CNN, can achieve 85.00% and 100% accuracies in identifying dysphonic voices with a sustained vowel (‘/a/’) and sentence samples, respectively. The proposed algorithm can also detect psychogenic and spasmodic dysphonia with high accuracy (i.e., 100%). To demonstrate the robustness of the proposed algorithm, the performances are evaluated with some popular machine learning algorithms. Finally, the performance comparisons of the proposed algorithm with some of the other related works are also presented in this Chapter to show its effectiveness.

6.1 Introduction

Automated voice pathology detection enables objective assessment and earlier diagnosis to assist clinicians with non-invasive diagnostics. One of the leading causes of voice pathology is the improper movement of vocal cords. The causes may include calluses and swelling on vocal cords, vocal cord paralysis, vocal cord shutting, and spasmodic dysphonia. This investigation addresses dysphonia, as it has been ranked by the American Speech-Language-Hearing Association (ASHA) as the most common voice pathology in the general population [140]. Generally, dysphonia refers to hoarse voices that occur suddenly or develop gradually. It may also change the pitch of the voice sound. Dysphonic

patients may produce rough, raspy, strained, weak, breathy, and gravelly voice sounds. They often complain of pain while speaking, singing, and projecting their voices. Inflamed vocal cords and other respiratory diseases may cause dysphonia [147]. Dysphonia is also correlated to a patient's psychological [216] and neurological [217] conditions. Many classifier-based algorithms have been recommended in the literature to detect voice pathology with reasonably high accuracy. Among them, ANNs and DNNs have recently drawn considerable attention from researchers in this field. It is shown in [11]-[12], [92], [126], [158]-[159], and [218] that deep learning algorithms can detect voice pathology with very high accuracy, provided appropriate voice features are excerpted.

This Chapter investigates a unique spectral feature called cochleagram [219]-[220] to identify dysphonic voices. Until now, cochleagram has not been used in pathological voice detection algorithms; however, it has been used in other applications that include the detection of acoustic events [221]-[222], audio surveillance [223], speech recognition [224], and speaker attitude detection [225]. The cochleagram is generated by an optimally designed gammatone filter bank. It uses a set of channels with distinct bandwidths mimicking the human cochlea. The bandwidths of these filters become increasingly more significant with higher frequencies. This exciting property assists the cochelagram in extracting more detailed spectral information in the voice samples compared to other popular spectral images like spectrogram and Mel-spectrogram.

Both vowels (‘/a/’) and sentence samples have been considered in this work. As stated earlier, a speaker can maintain a steady frequency and amplitude at a comfortable level during vowel sound generation. The vowel sound is an indicator of the phonation activity of the voice. On the other hand, sentence samples contain articulatory and other linguistic confounds that help diagnose the voice disorder due to neurogenic and other medical conditions. The significant contributions of this Chapter are as follows:

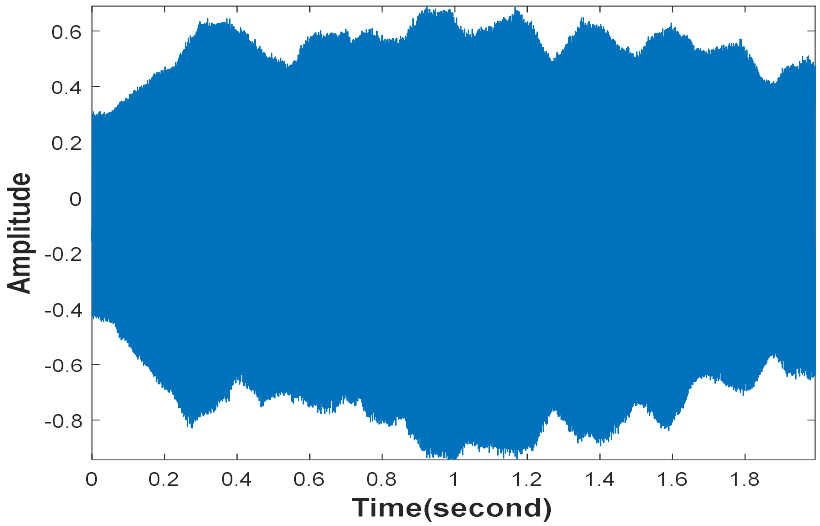
- Developing a novel pathological voice detection algorithm based on signal processing and a deep learning approach.
- Optimizing gammatone filters to extract the detailed spectral content of the audio signals.
- Computing the predominant audio spectral features and embedding them into cochleagram image form.

- Achieving a high classification accuracy by using a transfer learning approach, thus eliminating the limitation of sparse pathological data samples.
- Providing a detailed performance analysis of the proposed system in terms of the accuracy, precision, recall, F1 Score, negative predictive value (NPV), specificity, false negative rate (FNR), false detection rate (FDR), geometric mean (G-mean), and area under the curve (AUC).
- Optimizing the system performance with vowels (‘/a/’) and sentence samples for male and female patients to avoid gender biases.
- Verifying the achieved results with that of some popular machine learning algorithms.
- Applying the proposed algorithm to detect psychogenic and spasmodic dysphonia.
- Comparing the performances of the proposed algorithm with that of other related published works.

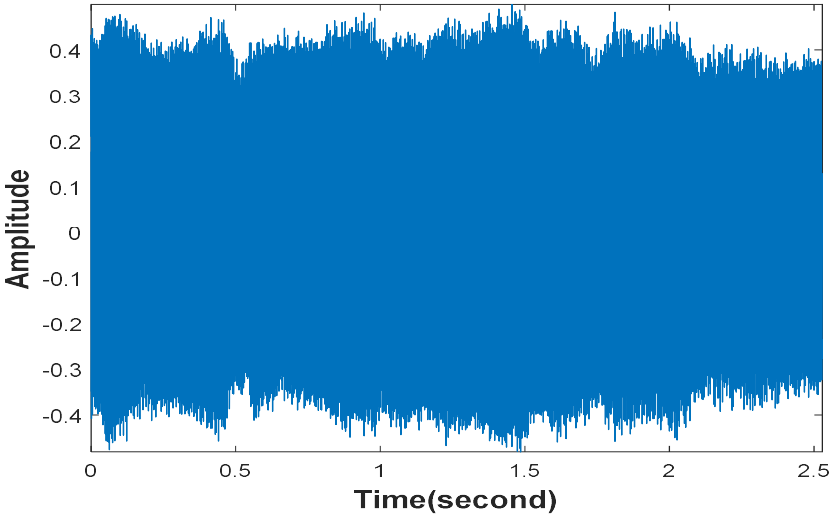
6.2 Materials and Methods

It is imperative to select suitable voice samples to ensure high accuracy in voice pathology detection. Existing pathological voice detection algorithms use voice samples that contain sustained vowel sounds, sentences, counting numbers, and running speeches. This investigation considers control (i.e., healthy) and dysphonic vowel ‘/a/’ and sentence samples from the SVD database [133]. Both vowel (with high tone) and sentence samples are considered as they have their own pros and cons. The sustained phonation of the vowel ‘/a/’ with a high tone is considered a lax vowel. The speaker can maintain a steady frequency and amplitude at a comfortable level [172] during the generation of this vowel. The vowel samples' limitation is that they only contain the voiced part. Primarily, vowel samples are used for vocal fold-related pathologies. In this work, the sentence speech samples “\Guten Morgen, wie geht es Ihnen?” (“Good morning, how are you?") have also been considered. The main advantage of these speech samples is that they contain both voiced and unvoiced components. Moreover, these samples contain articulatory and other linguistic confounds that often do not exist with the vowel samples. Fig. 6.1 shows the temporal plots for normal (i.e., healthy) and dysphonic vowel samples that are randomly collected from the SVD database. It is shown in the figure that the dysphonic voice sample

suffers from irregular distortion in both magnitude and phase compared to that of the healthy samples. Figure 6.2 shows the plot for normal (i.e., healthy) and dysphonic sentence samples from the SVD database as well. This figure shows that the dysphonic voice sample shows a more extended unvoiced phase compared to that of a healthy sample, in addition to irregular amplitude and phase variation.

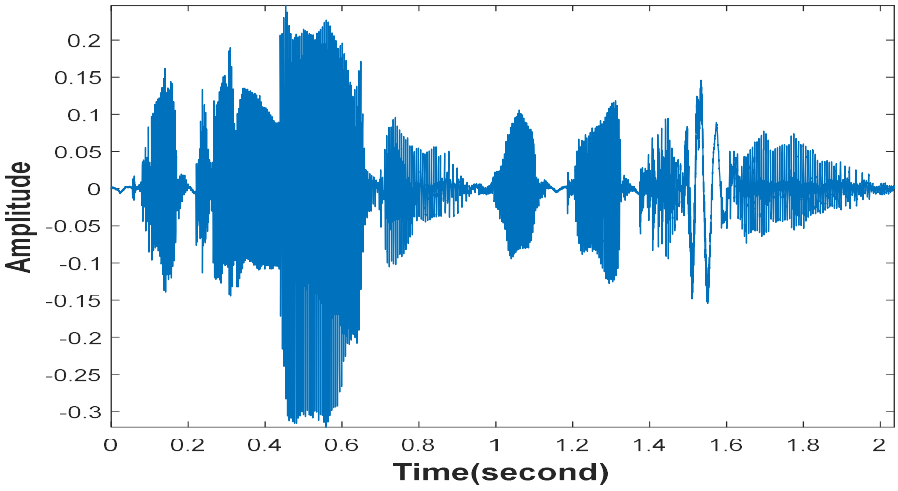


(a) Healthy sample

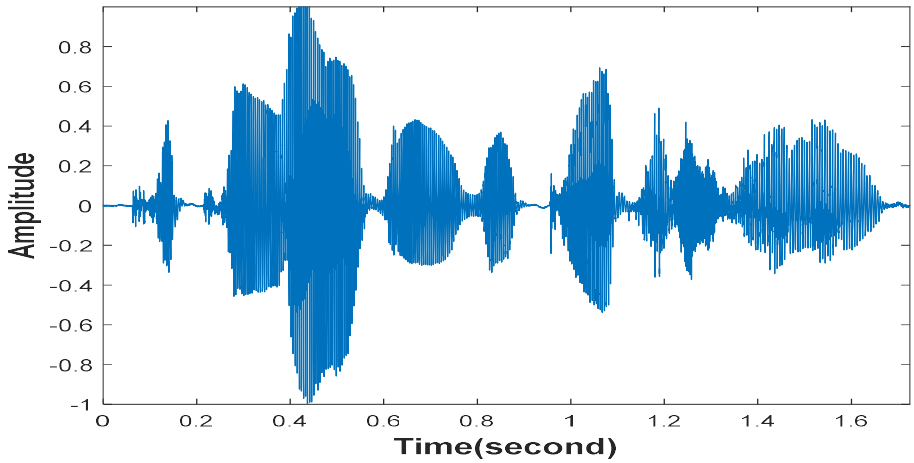


(b) Dysphonic sample

Figure 6.1. The (a) healthy and (b) dysphonic vowel (‘/a/’) samples.



(a) Healthy sample



(b) Dysphonic sample

Figure 6.2. The (a) healthy and (b) dysphonic speech samples of "\Guten Morgen, wie geht es Ihnen?\" ("Good morning, how are you?")

6.2.1 Male and Female Voice Samples

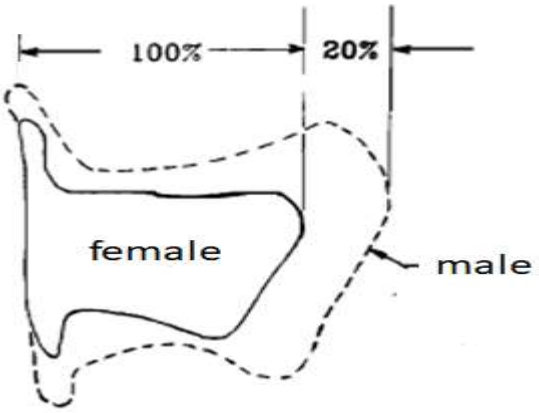
In this investigation, male and female voice samples are treated separately for the following reasons. The voice generation systems for males and females are structurally different [226]-[228]. For example, male and female larynges vary in size, vocal fold membranous length, the elasticity of vocal fold tissues, and pre-phonatory glottal shapes [226]-[228]. The anterior-posterior dimension of cartilage is approximately 20% longer, and the membranous vocal fold length is 60% longer in males [227], as shown in Fig. 6.3. The membranous vocal fold lengths of males and females also grow at a different rate [228].

For example, the membranous length of a male grows at a rate of 0.7 mm/year up to 20 years (approximately). On the other hand, the same grows at a rate of 0.4 mm/year for females. However, the membrane growth remains constant just after 20 years. When the larynx is fully developed, the vocal fold achieves a membranous length of about 10 mm and 16 mm for females and males, respectively. In [228], a relationship was driven between mean speaking fundamental frequency, F_0 and membranous vocal fold length, L_m. This relation is defined as $F_0 = \frac{1700}{L_m}$, where L_m is in mm. This formula predicts a fundamental frequency, $F_0 = 170$ Hz at L_m=10 mm for the adult female membranous length. The fundamental frequency is 106 Hz for the adult male membranous length of 16 mm. It is also shown in the same work that the female voice could be 25% more efficient than the male voice due to higher F_0. The authors concluded that a female's higher-pitched voice would appear more susceptible to tissue damage than a lower-pitched voice of a male. Considering the above issues, the proposed algorithm is trained and tested with the male and female voice samples separately to provide an unbiased classification accuracy.

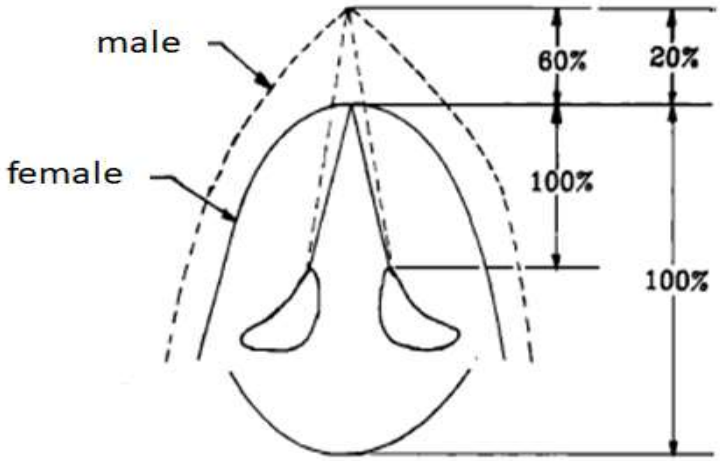
6.2.2 The Cochleagram

The cochleagram is an auditory image form that was initially developed to analyze sounds and music. Recently, it has been considered a more useful spectral image than traditional spectrogram and Mel-spectrogram. It has attracted researchers to implement various applications, including sound classification, voice recognition, voice activity detection, speaker attitude detection, and emotion detection. For example, the cochleagram has been used in [221] to classify 50 sound classes. In a similar work [222], the authors have used pseudo-color cochleagram images of sound signals for acoustic event recognition. It has also been used in an audio surveillance application [223]. To identify electronically disguised voices, a modified version of cochleagram has also been used [229]. A speech recognition system using deep CNN and cochleagram has been proposed in [224]. It is shown in [230] that the cocheagram can be used to detect a speaker's attitude too. Cochleagram has also been used in voice activity detection algorithms [231]-[233]. Multi-layer perceptron (MLP) and cochleagram have been used in [234] to detect vowels in natural language processing. A speaker identification algorithm under noisy conditions has been proposed in [235]. Cross-linguistic speech emotion detection is performed in [236].

Emotional data from English, Lithuanian, German, Spanish, Serbian, and Polish have been investigated in that work.



(a) Sagittal view



(b) Horizontal section

Figure 6. 3 Comparison of the larynx; (a) sagittal view and (b) horizontal section for males and females voices [227].

The cochleagram is produced by a gammatone filter bank. This filter bank performs spectral analysis and converts an acoustic wave into a multichannel representation that mimics the basilar membrane motion of the human cochlea. The impulse response, $h(t)$ of a gammatone filter, is defined by

$$h(t) = ct^{n-1}e^{-2\pi bt}cos(2\pi f_0 t + \varphi)u(t), \quad (6.1)$$

where c is a constant, n is the filter order, b is the temporal decay coefficient, f_0 is the carrier frequency, φ is the carrier phase, and $u(t)$ is a unit step function. The filter order n controls the relative shape of the filter, which becomes less skewed as n increases. The carrier phase, φ determines the relative position of the envelope. The parameter b determines the duration of the impulse response and hence determines the bandwidth of the gammatone filter. The Fourier transform of $h(t)$ is given by $H(f)$ and can be expressed as

$$H(f) = \frac{c}{2}(n-1)!\,(2\pi b)^{-n}\left[e^{j\varphi}\left(1 + j\frac{(f-f_0)}{b}\right)^{-n}\right] + \frac{c}{2}(n-1)!\,(2\pi b)^{-n}\left[e^{-j\varphi}\left(1 + j\frac{(f+f_0)}{b}\right)^{-n}\right]. \quad (6.2)$$

A complete derivation of the $H(f)$ can be found in [237]. This work uses a filter with order 4 for the following reasons. The gammatone filter with order 4 is very similar to that of the $roex$ function [238]-[239], and it represents the magnitude response of the human auditory filter [240]. An essential parameter of the gammatone filter is $\frac{f_0}{b}$. The selection of $\frac{f_0}{b} > 8$, makes the bandwidth proportional to b and becomes independent of f_0 [237]. In this work, $\frac{f_0}{b} = 9$ is selected. A filter bank consisting of 32 gammatone filters (with orders, $n = 4$ and $\frac{f_0}{b} = 9$*),* is used in this investigation. The magnitude spectrum of the gammatone filter bank used is shown in Fig. 6.4. The gain of the gammatone filters is adjusted according to the work presented in [241].

After filtering the voice signal with the gammatone filter, the energy in the windowed signal for each frequency component is added as follows: $C(m,n) = \sum_{i=0}^{N-1}\{|\hat{x}(m,i)w(i)|\}^2$, where $m = 1,2,\ldots.,M$, $\hat{x}(m,i)$ is the gammatone filtered signal, $w[i]$ is the ith window, $C(m,n)$ is the mth harmonic corresponding to the center frequency, f_{cm} for the nth frame, and M is the number of gammatone filters. The steps used to generate the cochleagram are shown in detail as Algorithm 1. The graphical representation of the cochleagram image for healthy and dysphonic patient produced by the proposed algorithm are shown in Figs. 6.5-6.6 for vowel and sentence samples, respectively. The cochealgram of the healthy vowel sample shows that the prominent frequencies range

between 12.5 and 17.5 Hz (marked by the red colour). On the other hand, the significant frequencies are scattered in the different bands of frequencies for dysphonic voice. For example, the dominant frequencies are in the range of 7.5-11 Hz (marked by the red colour) for the dysphonic voice. In addition, some significant frequencies are concentrated in the 13-15 Hz (marked by red and yellow colour) range. The cochleagrams of healthy and dysphonic speech samples also demonstrate differences in spectral contents, as shown in Fig. 6.6.

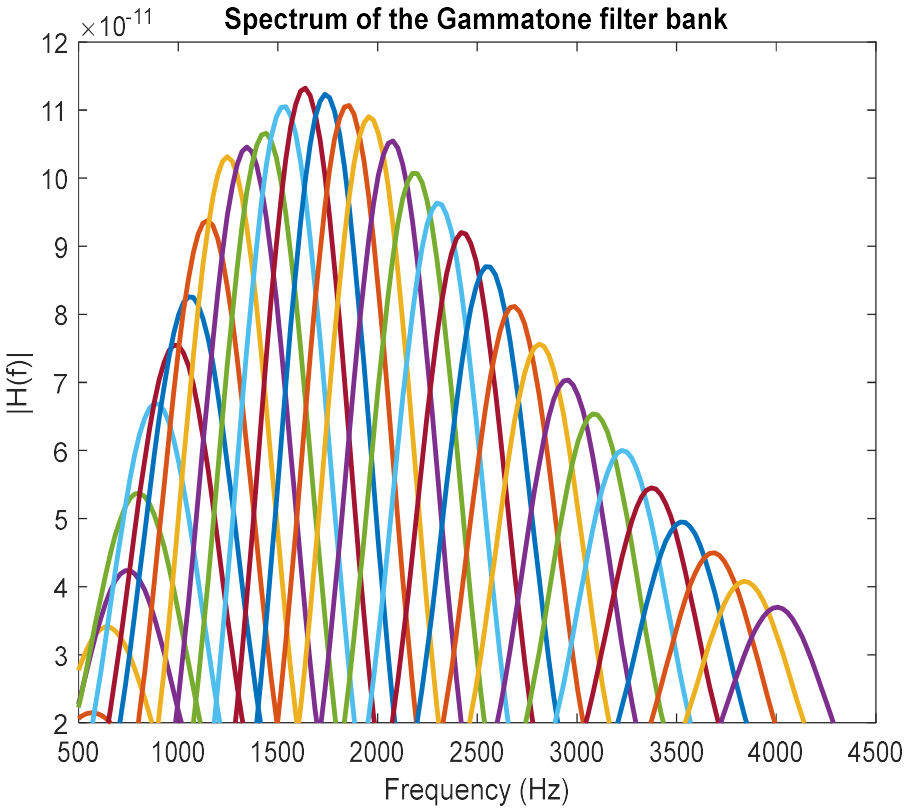
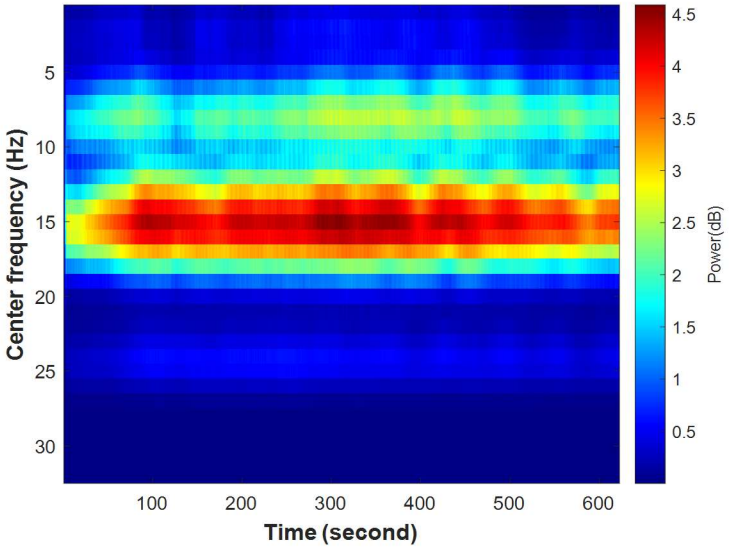


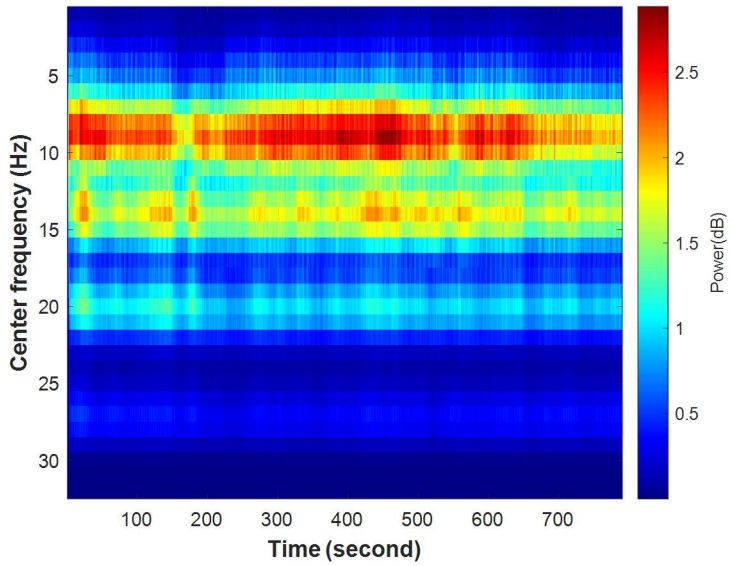
Figure 6.4. The spectrum of the Gammatone filter bank.

The cochleagram images of dysphonic and healthy voice samples are applied to the input of a CNN for classification purposes. The CNNs have been widely used in image classification tasks because of their flexible nature. The CNN, introduced in [242], shows its capability to recognize the patterns in an image irrespective of their orientation. One of the major limitations of the deep CNN models is that they deal with significant parameters

and hence require a special GPU to handle them. In addition, the CNN models require a large dataset for training. However, the publicly available voice databases (including the SVD database) provide a limited dataset for healthy, especially for pathological samples. To overcome this limitation, a transfer learning approach is adopted in this work. In this approach, the learned weights of pre-trained CNN models are used for identifying the lower-level features, and consequently, the network is trained to learn the upper-level features from the given data.



(a) Healthy sample

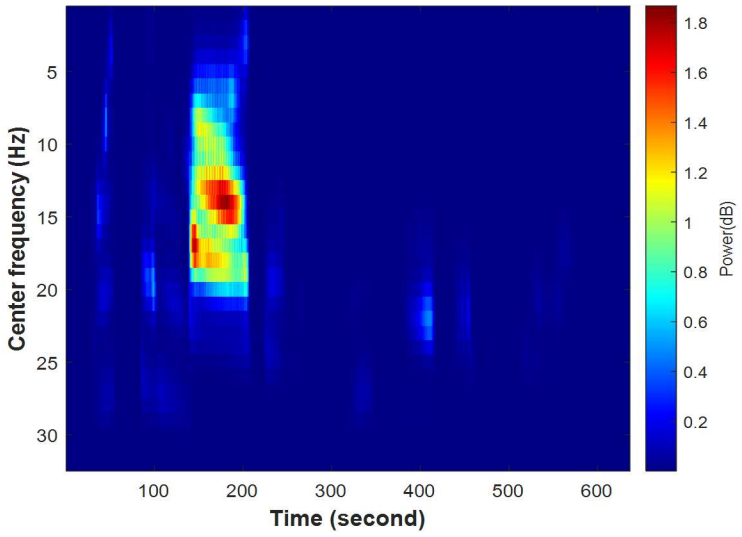


(b) Dysphonic sample

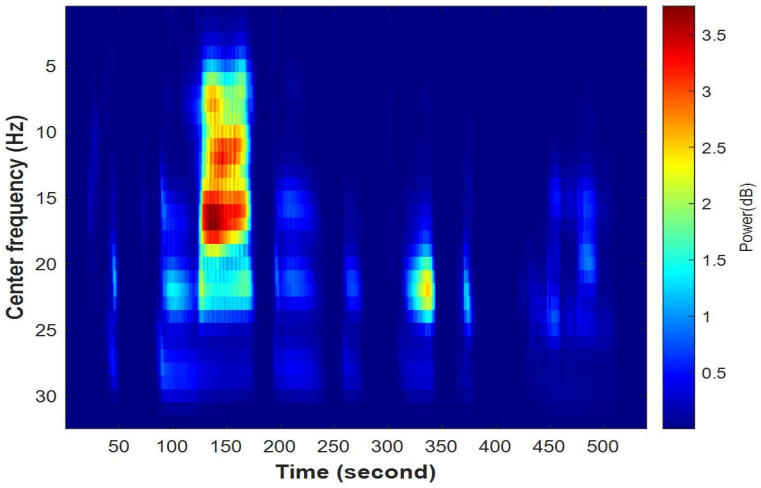
Figure 6.5 The (a) healthy and (b) dysphonic cochleagram images for vowel (‘/a/’) samples.

Algorithm 1 : Generating the cochleagram of voice signals

Set the lowest value of the frequency $\mathbf{f_{min} = 50}$;
Set the highest value of the frequency $\mathbf{f_{max} = 400}$;
Set the sampling frequency rate $\mathbf{F_s = 50000}$;
$\mathbf{N \leftarrow 8}$; /* specify the filter order */
$\mathbf{L \leftarrow 2048}$; /* specify the filter length */
$\mathbf{C \leftarrow 32}$; /* specify the number of channels */
$\emptyset \leftarrow \mathbf{0}$; /* specify the phase */
/* calculate the lowest value of the equivalent rectangular bandwidth */
$\mathbf{ERB_{min} = 21.4\left[\log_{10}\left\{\frac{4.37f_{min}}{1000} + 1\right\}\right]}$
/* calculate the maximum value of the equivalent rectangular bandwidth */
$\mathbf{ERB_{max} = 21.4\left[\log_{10}\left\{\frac{4.37f_{max}}{100} + 1\right\}\right]}$
/* calculate the equivalent rectangular channel bandwidth */
$\mathbf{ERB_i = \frac{ERB_{max} - ERB_{min}}{C}}$
/* compute the center frequencies of each filter */
$\mathbf{f_{c,i} = \frac{100\left(10^{\frac{ERB_i}{21.4}} - 1\right)}{4.37}}$
/* calculate the filter decay rate */
$\mathbf{b = \frac{110f_{ci}}{1000} + 25.16}$
/* specify the gammtone filters impulse response */
$\mathbf{h_i(t) = ct^{n-1}e^{-2\pi bt}\cos(2\pi f_{ci}t + \varphi)u(t)}$
Adjust gain adjustment of the filter transfer function depending on the loudness level
Set the path to the directory and read the voice samples
/* Count the number of files and store it */
$\mathbf{N \leftarrow Count}$
$\mathbf{i \leftarrow 1}$
while $\mathbf{i \leq N}$ **do**
    Load the voice sample file and store the data into a variable, **X**
    /* convert *X* into vector */
    $\vec{\mathbf{V}} \leftarrow \mathbf{X}$
    Filter the vectored sound signal, $\vec{\mathbf{V}}$ by the gammatone filter bank
    Store the filter output into a variable, $\vec{\mathbf{Y}}$
    Divide vector $\vec{\mathbf{V}}$ into frames, $\vec{\mathbf{F}}$
    Calculate the energy of each frame, **E**
    Convert **E** into image form and plot on a time-frequency scale
end while



(a) Healthy sample



(b) Dysphonic sample

Figure 6.6. The (a) healthy and (b) dysphonic cochleagram images for speech samples.

A pre-trained deep CNN network model called VGG16 [243] is used for this investigation. The VGG16 is already trained on large datasets (i.e., 20,000 categories of images) and is available as pre-packaged with the Keras. The VGG16 used in this work consists of a stack of convolutional layers followed by three fully connected layers. The proposed system model is shown in Fig. 6.7. The first two fully connected networks have 4096 channels each. The final layer is the Softmax layer. All hidden layers are equipped

with the activation function, ReLU. There are 134,268,738 parameters in the network. Among those parameters, 8194 parameters are trained, and the other 134,260,544 parameters are left untrained. The rescaled version of the RGB images of cochleagrams with a fixed size of 224×224 is applied to the input of convolutional layer one. These images are then passed through the rest of the convolutional layers, where the filters are used with a very small receptive field. Spatial pooling is carried out by five max-pooling layers. The max-pooling is performed over a 2×2 pixel window with stride 2. During training, the learned weights from these pre-trained models are used by freezing the model's upper layers, and the lower fully connected layers are trained on the cochleagram images. The detailed classification steps are explained in Algorithm 2. In this work, image classification and predictions are performed in Google Colaboratory (also called Colab) with Python 3 Google Compute Engine. The audio file processing and feature extraction (to generate cochleagram) algorithm were implemented in MATLAB 2020.

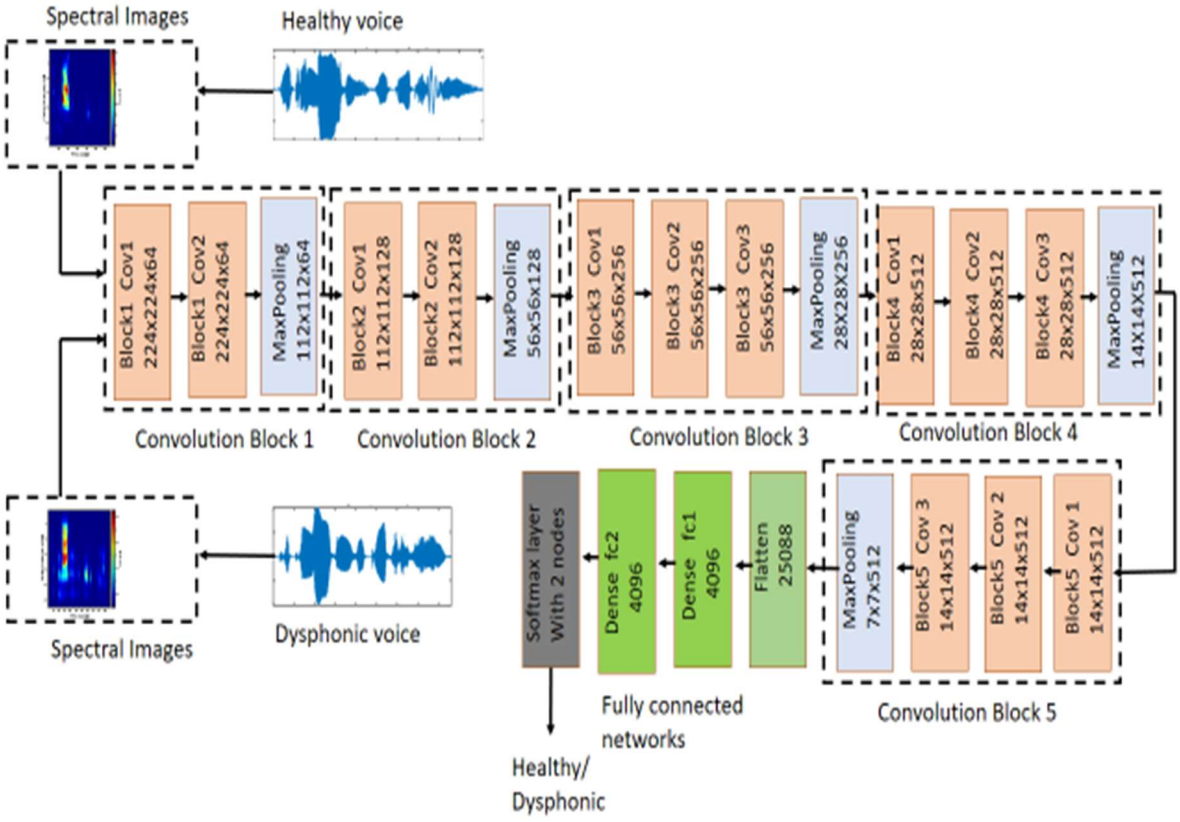


Figure 6.7. The System model employing VGG16.

Algorithm 2: Steps to classify the voice samples

1: **Load** images for training, *I*
2: **Load** images for testing, *T*
3: Rescale the images to (224, 224, 3)
4: Normalize the pixel values of the images between 0 and 1
5: **Load** the pre-trained VGG16 model
6: Freeze the base network
7: Flatten the network
8: Add two dense network layers on top of the base network
9: Extract the feature map, ***F*** from the training images, ***T*** using pre-trained VGG16
10: Split the images, ***I*** into training and validation set in the ratio of 7:3
11. **Set** the epoch, ***N***
11. **Set** the counter i
12. **do while** $i < N$
11: **Select** the initial hyper parameter values (e.g., learning rate, batch size, etc)
12: Train the classifier using the training dataset
13: The validation set is used to evaluate VGG16 performance during the training phase
14: **end**
15: Choose the best candidate typical with a minimum validation error rate
16: Generate prediction scores, *P* based on testing samples
17: Use predict the score, ***P*** to classify samples as healthy or dysphonic
18: Through utilizing the test samples, the best-trained model is identified

6.3 Simulation Results

The proposed algorithm classifies the voice samples into healthy and dysphonic. Classification decisions for pathological voices are justified in the context of medical diagnosis and statistical measures of accuracy and validity. The performances of the proposed system are evaluated with the commonly accepted measures of accuracy, precision, recall, F1 Score, NPV, and specificity. Additionally, false-negative rate (FNR), false discovery rate (FDR), and geometric mean (G-mean), as described in the following equations (6.3)-(6.5) are also evaluated in terms of TP, TN, FP, and FN.

False Negative Rate (FNR) is the proportion of incorrectly classified observations per true class. It is expressed as

$$\mathrm{FNR} = \mathrm{FN}/(\mathrm{FN} + \mathrm{TP}) \tag{6.3}$$

False Discovery Rate (FDR) is the proportion of incorrectly classified observations per predicted class. For pathological sample, it is expressed as

$$\text{FDR} = \text{FP}/(\text{FP} + \text{TP}) \tag{6.4}$$

The geometric mean (G-mean) metric is calculated as the geometric mean of the sensitivity and specificity metrics. It is defined by

$$\text{G–mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \tag{6.5}$$

The performance of the proposed algorithm is investigated using a cochleagram of 100 healthy voice samples and 100 dysphonic voice samples for both sustained vowel (‘/a/’) and phrase/sentences, respectively. The hyperparameters are adjusted to maximize the training accuracy. A list of these hyperparameters and their optimum values is provided in Table 6.1. The performances of the system in terms of the evaluation parameters, namely TP, TN, FP, and FN, are listed in Table 6.2. The table shows that the highest TP (90) and TN (100) were achieved with the female phrase samples for dysphonic voices.

The performance metrics in terms of accuracy, precision, recall, F1-Score, NPV, specificity, FNR, FDR, G-mean, and AUC are listed in Table 6.3. This table shows that the highest accuracy of 95% is achieved with the cochleagram of female phrase samples. The precision and specificity are both 100% for the same voice samples. The other performance parameters, recall, F1 Score, NPV, G-mean, FNR, FDR, and AUC, are also best for female phrase samples to identify dysphonic voices.

Table 6.1 The best trained VGG16 model parameters

Hyperparameters	
Optimization method	RMSprop
Training mode	auto
Patience (early stopping)	10
Dropout	25%
Batch size	10
Learning rate	0.0001
Epoch	30
Minimum detectable loss	0.00001
Learning reduction rate	0.1

Table 6.2 Performance parameters for dysphonic voice with VGG16

Gender	Phoneme	TP	TN	FP	FN
Female	vowel	80	90	10	20
	phrase	90	100	0	10
Male	vowel	75	87.5	12.5	25
	phrase	75	75	25	25

Table 6.3 Performances of VGG16 for dysphonic voice detection

Measures	**Cochleagram**			
	Female		Male	
	Vowel	Phrase	Vowel	Phrase
Accuracy (%)	85.00	95.00	81.25	75.00
Precision/PPV (%)	88.89	100.00	85.71	75.00
Recall/Sn/TPR (%)	80.00	90.00	75.00	75.00
F1 Score (%)	84.21	94.74	80.00	75.00
NPV (%)	81.82	90.91	77.78	75.00
Specificity/TNR (%)	90.00	100.00	87.50	75.00
FNR (%)	20.00	10.00	25.00	25.00
FDR (%)	11.11	0.00	14.29	25.00
G-mean (%)	84.85	94.87	81.01	75.00
AUC	0.85	1	0.81	0.75

The proposed algorithm achieves accuracies of 81.25% and 75% with the vowel and phrase samples, respectively, for male subjects. The performance measures in terms of precision, recall, F1Score, NPV, specificity, FNR, FDR, G-mean, and AUC are all better for vowel samples in comparison with the phrase for male subjects, as observed in Table 6.3. Based on the observation mentioned above, it can be concluded that female voice pathology is more accessible to detect than its male counterpart. It supports that female patients are more susceptible to voice pathology than male voices due to having higher fundamental frequency; hence, voice pathologies are more accessible to detect in females than in males. The simulation results also concluded that the phrase voice samples contain more articulatory and other linguistic confounds compared to vowel samples and hence are more helpful in detecting dysphonic voices for female patients.

The proposed algorithm is also tested on the phrase samples of males and females suffering from psychogenic dysphonia [216]. Psychogenic dysphonia refers to voice pathology without the sufficient structural or neurological disorder. This pathology has been correlated to psychological processes, including anxiety, depression, conversion

reaction, and personality disorder. The results in terms of performance measures are presented in Table 6.4. This table shows that the proposed algorithm can detect psychogenic dysphonia with very high accuracy. The achieved accuracies are 100% with the female phrase samples and 80% with the male phrase samples.

Table 6.4 Performances of VGG16 for detection of psychogenic dysphonia with phrase samples

Measures	Cochleagram	
	Female	Male
Accuracy (%)	100.00	80.00
Precision/PPV (%)	100.00	80.00
Recall/Sn/TPR	100.00	80.00
F1 Score (%)	100.00	80.00
NPV (%)	100.00	80.00
Specificity/TNR	100.00	80.00
FNR (%)	0.00	20.00
FDR (%)	0.00	20.00
G-mean	1	0.8
AUC	1	0.8

In this work, the classification measures are experimented with phrase voice samples of spasmodic dysphonia also. This pathology is caused by abnormal functioning in an area of the brain called the basal ganglia [217]. The main function of the basal ganglia is to coordinate the movements of muscles throughout the body. Recent research has found abnormalities in other regions of the brain associated with a particular type of voice pathology called spasmodic dysphonia. This area of the brain includes the cerebral cortex that controls commands to muscles and coordinates these commands with incoming sensory information. The experimental results with the cochleagram of the spasmodic dysphonia samples are presented in Table 6.5. This table shows that the proposed algorithm can detect spasmodic dysphonia with 100% and 70% accuracy in the female and male samples, respectively. The precision, recall, F1 Score, NPV, specificity, and G-meal are 100% for the female samples. On the other hand, these values are 66.67%, 80%, 72.73%, 75%, 60%, and 69.28%, respectively, for the male phrase samples. The FNR and FDR are both 0% for female spasmodic dysphonic phrase samples. The same measures are 20% and

33.33% for male spasmodic samples, respectively. The AUCs are 1.0 and 0.7 for female and male samples respectively. Again, a similar conclusion can be made here: the female voice samples provide much better performance metrics than their male counterparts, even for spasmodic dysphonia.

Table 6.5 Performances of VGG16 for
detection of spasmodic dysphonia with phrase samples

Measures	Cochleagram	
	Female	Male
Accuracy (%)	100.00	70.00
Precision/PPV (%)	100.00	66.67
Recall/Sn/TPR	100.00	80.00
F1 Score (%)	100.00	72.73
NPV (%)	100.00	75.00
Specificity/TNR	100.00	60.00
FNR (%)	0.00	20.00
FDR (%)	0.00	33.33
G-mean	1	0.7
AUC	1	0.7

To validate the results achieved with the pre-trained VGG16 model, the experimental simulations are repeated with the Statistics and Machine Learning (SML) Toolbox of MATLAB 2020 using cochleagram images of the same voice samples. In this investigation, several machine learning algorithms have been considered. The results are presented in Table 6.6. The best performance metrics for the female dysphonic samples are achieved with the quadratic SVM using phrase dysphonic voice samples. The data presented in this table depicts similar conclusions that were achieved with the VGG16. For example, the best accuracy of 81.8% is achieved with the phrase voice samples for female subjects. The other performance parameters, namely, precision, recall, F1 Score, NPV, specificity, FNR, FDR, and G-mean are 80.00%, 87.80%, 83.72%, 84.38%, 75.00%, 12.20%, 20.00%, and 81.15%, respectively. These values are better compared to other investigated machine learning algorithms, but the specificity. The highest specificity of 77.78% is achieved with the ensemble subspace discrimination algorithm with vowel samples.

The machine learning algorithms are also used for the classifications of male voice samples. Most of the best performances can be achieved with the ensemble subspace kNN for the phrase voice samples. The highest accuracy, recall, F1-score, NPV, AUC, and G-mean are achieved with this classifier. All the measures are shown in Table 6.8. All the performance measures are better for female phrase samples as compared with male phrase samples, as visualized in Tables 6.6-6.7.

Table 6.6 Performances of machine learning algorithms for detection of female dysphonic voice

Parameters	**Medium Gaussian SVM**	**Quadratic SVM**	**Ensemble subspace discriminant**	**Ensemble subspace discriminant**	**Ensemble subspace kNN**	**Cosine kNN**
	vowel	**phrase**	**vowel**	**phrase**	**vowel**	**phrase**
Accuracy (%)	74.40	**81.80**	74.40	76.60	73.30	70.10
Precision (%)	73.91	**80.00**	76.19	50.85	74.42	68.75
Recall/Sn/ TPR (%)	75.56	**87.80**	71.11	73.17	71.11	80.49
F1 Score (%)	74.73	**83.72**	73.56	60.00	72.73	74.16
NPV (%)	75.00	**84.38**	72.92	38.89	72.34	72.41
Specificity/ TNR (%)	73.33	75.00	**77.78**	19.44	75.56	58.33
FNR (%)	24.44	**12.20**	28.89	26.83	28.89	19.51
FDR (%)	26.09	**20.00**	23.81	49.15	25.58	27.59
G-mean	0.7444	**0.8115**	0.7437	0.3772	0.7330	0.6852
AUC	0.79	**0.86**	0.78	0.79	0.79	0.78

Male psychogenic and spasmodic dysphonia can be detected using machine learning algorithms. The best results are achieved with linear discriminant algorithms. All the performance metrics are shown in Table 6.8.

Table 6.7 Performances of machine learning algorithms for the detection of male dysphonic voice

Parameters	Linear SVM	Linear SVM	Ensemble subspace discrimination	Ensemble subspace kNN	Weighted kNN	Fine kNN
	vowel	phrase	vowel	phrase	vowel	phrase
Accuracy (%)	67.10	70.30	65.90	71.60	62.20	70.30
Precision (%)	68.42	72.73	67.57	70.00	60.42	72.73
Recall/Sn/TPR (%)	63.41	64.86	60.98	75.68	70.73	64.86
F1 Score (%)	65.82	68.57	64.10	72.73	65.17	68.57
NPV (%)	65.91	68.29	64.44	73.53	64.71	68.29
Specificity/TNR (%)	70.73	75.68	70.73	67.57	53.66	75.68
FNR (%)	36.59	35.14	39.02	24.32	29.27	35.14
FDR (%)	31.58	27.27	32.43	30.00	39.58	27.27
G-mean	0.6697	0.7006	0.6567	0.7151	0.6161	0.7006
AUC	0.7	0.73	0.73	0.73	0.61	0.7

Table 6.8 Performances of machine learning algorithms for male psychogenic and spasmodic dysphonia

Parameters	Linear Discriminant (Psychogenic)	Linear Discriminant (Spasmodic)
	Phrase	Phrase
Accuracy (%)	70.00	77.50
Precision (%)	80.00	76.19
Recall/Sn/TPR (%)	53.33	80.00
F1 Score (%)	64.00	78.05
NPV (%)	65.00	78.95
Specificity/TNR (%)	86.67	75.00
FNR (%)	46.67	20.00
FDR (%)	20.00	23.81
G-mean	0.6799	0.7746
AUC	0.7	0.78

The performances of the machine learning algorithms for detecting psychogenic and spasmodic dysphonia are presented in Table 6.9 for female samples. The data presented in Table 6.9 shows that machine learning algorithms, including medium Gaussian SVM,

cubic SVM, quadratic SVM, and ensemble bagged trees, provide the maximum accuracy of 100%, which is much higher than that of the male phrase samples (Table 6.8). These results also confirm the 100% accuracy obtained by the VGG16 pre-trained network.

Finally, the performance of the proposed model is compared with that of some other approaches in the literature for the diagnosis of dysphonia, and the comparison is listed in Table 6.10. The comparison shows that the proposed system achieves an accuracy of 100% with the female speech samples, which is higher than that of the works presented in [244]-[247]. This algorithm also can identify dysphonic female voices from vowels (‘/a/’) samples but with a lower accuracy (i.e., 85%). However, the accuracy is higher than the algorithm presented in [247]. This accuracy is also very close to the accuracy provided by the work presented in [245]. This table also shows that the proposed algorithm can detect psychogenic and spasmodic dysphonia. These pathologies are hard to identify by traditional invasive methods as they are neurogenic. However, the results show that the proposed algorithm can detect them for female samples with an accuracy of 100%, which is also higher than the accuracies provided by the algorithms presented in [244]-[247].

6.4 Conclusion

This Chapter has proposed a robust, non-invasive, and automated voice pathology detection system to detect dysphonia pathology. The proposed system has been optimized for sustained vowel (‘/a/’) and speech samples. Most of the works in literatue considered vowel samples. Also, very few investigations were established with speech samples. The proposed system has achieved an accuracy of 100% for speech samples. The achieved higher accuracy for the speech sample is well justified, as speech includes phonation, articulation, and prosody of voice. In clinical settings, the diagnosis of voice pathology is limited either through perceptual judgment or objective invasive assessment. Dysphonia is a perceptual quality of voice that indicates some negative changes have occurred in the voice generation system. The presented perceptual analysis through signal processing and a deep learning-based approach can be a promising tool for the diagnosis of voice disorder. In practice, the perceptual analysis is costly in time and involves human resources. The proposed automated system can correlate the clinical findings and monitor the treatment progress for dysphonic voice. Also, the extracted predominant spectral feature i.e. cochleagram is promising to quantify the healthy and dysphonic voice. This system

overcomes the computational burden with a pre-trained VGG16 network using a transfer learning approach. The limitation of a small dataset for dysphonic samples has been eliminated in this method.

Table 6.9 Performances of machine learning algorithms for female psychogenic and spasmodic dysphonia

Female Psychogenic Dysphonia				
Parameters	**Medium Gaussian SVM**	**Cubic SVM**	**Quadratic SVM**	**Ensemble Bagged Trees**
	vowel	vowel	vowel	vowel
Accuracy (%)	100.00	100.00	100.00	100.00
Precision (%)	100.00	100.00	100.00	100.00
Recall/Sn/TPR (%)	100.00	100.00	100.00	100.00
F1 Score (%)	100.00	100.00	100.00	100.00
NPV (%)	100.00	100.00	100.00	100.00
Specificity/TNR (%)	100.00	100.00	100.00	100.00
FNR (%)	0.00	0.00	0.00	0.00
FDR (%)	0.00	0.00	0.00	0.00
G-mean	1	1	1	1
AUC	1	1	1	1
Female Spasmodic Dysphonia				
Parameters	**Linear discriminant**	**Linear SVM**	**Logistic regression**	**Cosine kNN**
	vowel	vowel	vowel	vowel
Accuracy (%)	100.00	100.00	100.00	100.00
Precision (%)	100.00	100.00	100.00	100.00
Recall/Sn/TPR (%)	100.00	100.00	100.00	100.00
F1 Score (%)	100.00	100.00	100.00	100.00
NPV (%)	100.00	100.00	100.00	100.00
Specificity/TNR (%)	100.00	100.00	100.00	100.00
FNR (%)	0.00	0.00	0.00	0.00
FDR (%)	0.00	0.00	0.00	0.00
G-mean	1	1	1	1
AUC	1	1	1	1

In this work, it was also demonstrated that the proposed algorithm could detect dysphonic voices that are related to neurogenic conditions. Psychogenic dysphonia and spasmodic dysphonia are examples of such types of voice diseases. The results have shown that the proposed algorithm can detect these two pathologies in female patients with an accuracy of 100%, provided phrase samples are used.

Table 6.10 Performance comparisons of the proposed system with other related works

Research works	Phonemes	Pathological Condition	Features	Tools	Accuracy
Little, A.M. [244]	Sustained vowel	Dysphonia	Pitch Period Entropy (PPE)	Kernel SVM	91.40%
Zuzana Dankovi˘cová [245]	vowels /*a*/, /*i*/, and /*u*/	Dysphonia	shimmer, jitter, MFCC, spectral roll-off, spectral flux, etc	SVM, random forests classifier (RFC), kNN	86.20%
Shanshan Yang [246]	Sustained vowel	Dysphonia	F0, MDVP: Jitter (%), DFA, and Spread2	MAP	91.80%
João Paulo Teixeiraa [248]	vowel '/a/'	Dysphonia	Jitter, shimmer, HNR	ANN	Female: 100% Male: 90%
L. Verde [247]	Sustained vowel '/a/'	Dysphonia	F0, jitter, shimmer, and HNR	Tree regression	82.60%
Proposed Method	Sustained vowel: '/a/'	Dysphonia	Cochleagram	VGG16, Machine learning	**Female: 85%** Male: 81.25%
	Phrase/ Sentence: "\Good morning, How are you?\"	Dysphonia	Cochleagram	VGG16, Machine learning	**Female : 95%** Male: 75%
	Phrase/ Sentence: "\Good morning, How are you?\"	Psychogenic Dysphonia	Cochleagram	VGG16, Machine learning	**Female: 100%** Male: 70%
	Phrase/ Sentence: "\Good morning, How are you?\"	Spasmodic Dysphonia	Cochleagram	VGG16, Machine learning	**Female: 100%** Male: 77.5%

Some popular machine learning algorithms have also been experimented to verify the accuracy achieved by the proposed algorithm. The detection accuracy achieved with

these machine learning algorithms also conforms to the similar pattern obtained by the proposed algorithm. These machine learning algorithms also provided higher accuracy in detecting pathology in female samples compared to male samples. In addition, these machine learning algorithms also achieved higher accuracy with the phrase samples compared to vowel samples.

The proposed system could be extended to evaluate other voice disorders in the future. The proposed algorithm can also be extended to detect the progression levels of dysphonic patients. Also, investigations of classification network performances with other transfer learning-based approaches are left as future work.

CHAPTER 7
CONCLUSIONS AND FUTURE WORKS

7.1 Conclusions

Objective and noninvasive assessment techniques for voice disorder detection were addressed in this dissertation, considering four strategies. Comprehensive survey work on voice disability detection algorithms has also been provided in Chapter 2. The first part of this Chapter presented background information, including causes of voice disability, current procedures and practices, voice features, and classifiers. Graphical analyses have been presented from experimental simulations of a healthy, and an ASD child's voice signal to describe discriminative voice features. The issues and challenges related to the selection of voice features and classifier algorithms have been addressed at the end of that Chapter.

The successful detection of physiologic and neurologic voice pathologies depends on selecting appropriate voice samples (vowel/sentence) that correlate with the clinical aspects as demonstrated in this research. Also, choosing proper voice signals: speech or EGG, to identify and classify pathological voices is crucial, manifested in this dissertation. An algorithm was developed in Chapter 3 to diagnose and classify three major vocal fold pathologies: dysphonia, laryngitis, and vocal fold cysts from sustained vowel ('/a/') samples of both speech and EGG signals. The merit of the designed CNN-based algorithm is that it can process raw speech and EGG samples to preserve pathological information within the dataset. Binary detection accuracy (healthy voice/diseased) is excellent with speech signals. The multi-classification performance strongly depends on pathology attributes, as revealed in this study. The EGG signal that can adequately mimic the vibratory pattern of vocal folds exhibits better performance for categorizing vocal fold diseases. The designed algorithm can work with small datasets, thus reducing computation burden, and is clinically viable.

A novel voice pathology detection model was developed considering the biological process of speech perception in Chapter 4. A gamatone filterbank is designed following the signal processing steps of the cochlear implant model to extract discriminative information from speech (sentence) samples of healthy and laryngeal voice disorder patients. The critical center frequencies of those filters were selected to mimic the human

cochlear vibration patterns caused by audio signals. The superiority of the designed model with the conventional approach is that it eliminates the need for feature extraction from the speech samples. The performance matrix has shown significant improvement considering optimized gammatone filterbank instead of conventional bandpass filters (about 1.1% improvement in F1 Scores). The additional advantages of the gammatone filters as mentioned are that they (a) provide an appropriate "pseudo-resonant" frequency transfer function, (b) demonstrate a simple impulse response, and (c) support efficient hardware implementation [25].

A respiratory disease detection algorithm considering COVID-19 and healthy coughing sound samples was implemented to aid noninvasive diagnostics in Chapter 5. Predominant acoustic feature vectors in three domains: time, frequency, and mixed, are extracted from coughing sounds to design this model. The significant accuracy (97.5%) has been achieved with a feature vector in the frequency domain and the design of a DNN algorithm with a dropout strategy. This system can work as an alternative to painful PCR tests for preliminary assessment of COVID-19. Also, this noninvasive technique can prevent contact tracing or the spreading of the disease.

The fourth pathological voice detection algorithm was designed considering an auditory perceptual analysis of voice signals in Chapter 6. Cochleagrams were generated from healthy and dysphonic voice samples. Using a transfer learning approach, this proposed system overcomes the computational burden with a pre-trained VGG16 network. The limitation of a small dataset for dysphonic voice samples was eliminated in this method. The proposed approach was optimized for both sustained vowel (‘/a/’) and speech samples. The male and female voices were treated separately to develop this algorithm as they have physiological differences in their voice generation systems. The highest accuracy achieved with the proposed system is 100% for speech samples of female voices.

7.2 Future works and research applicability

Multimodal approaches based on the integration of clinical, neurophysiological, neuropsychological, and imaging measures to promote the current comprehension of the pathological voice are essential. Methodological advancement to consider nonlinear dynamicity of speech samples is necessary. The objective assessment of the human voice through spectral analysis and artificial intelligence would open new opportunities for the

understanding and follow-up of neurologic voice disorders in line with telemedicine approaches. As a future work, more focus can be given to identify the other neurological and respiratory diseases that strongly correlate with the voice generation system's impairment. Processing the voice samples of the alcoholic anonymous group, having the highest risk of cancer in the head and neck region, can be investigated to identify signature changes of voice at the early stage of the disease. Thus, minimizing the risk of spreading the disease through early diagnostics. The objective and noninvasive voice signal evaluation of pre- and post-operative patients with established vocal issues can be assessed to understand the state of recovery. The aspirated voice analysis of laryngeal cancer patients can objectively correlate the stages of cancer. Ninety percent (90%) of lung cancer patients' voice is dysphonic; identification of the prevalence of dysphonia can be a biomarker for early diagnosis of lung cancer.

REFERENCES

[1] "Voice Disorders," American Speech-Language-Hearing, 30 March 2020. [Online]. Available: https://www.asha.org/ practice-portal/clinical-topics/voice-disorders/. [Accessed 01 September 2022].

[2] T. E. Quateri, Discrete-Time Speech Signal Processing: Principles and Practices, Upper Saddle River: Prentice Hall, 2001.

[3] D. Zhang and K. Wu, Pathological Voice Analysis, Singapore: Springer, 2020.

[4] S. Collins, "Direct and Indirect Laryngoscopy: Equipment and Techniques," *Respiratory Care,* vol. 59, p. 850–864, 2014.

[5] D. Mehta and R. Hillman, "Current role of stroboscopy in laryngeal imaging," *Current Opinion Otolaryngology - Head Neck Surgery,* vol. 20, p. 429–436, 2012.

[6] Y. Heman-Ackah, S. Mandel, R. Manon-Espaillat, M. Abaza and R. Sataloff, "Laryngeal Electromyography Otolaryngology Clinic," *Otolaryngologic Clinics of North America,* vol. 40, p. 1003–1023, 2007.

[7] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, K. Malki, T. Mesallam and M. Ibrahim, "Voice Pathology Detection and Classification using Auto-correlation and entropy features in Different Frequency," *IEEE Access,* vol. 6, p. 6961–6974, 2017.

[8] D. Taib, M. Tarique and R. Islam, "Voice Features Analysis for Early Detection of Voice Disability in Children," in *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology*, Louisville, 2018.

[9] S. Hegde, S. Shetty, S. Rai and T. Dodderi, "A Survey on Machine Learning Approaches for Automatic Detection of Voice Disorder," *Journal of Voice,* vol. 33, p. 947.E11–947.E33, 2019.

[10] R. Islam and M. Tarique, "Classifier Based Early Detection of Pathological Voice," in *Proceedings of the International Symposium on Signal Processing and Information Technology*, Ajman, 2019.

[11] R. Islam, E. Abdel-Raheem and M. Tarique, "A study of using cough sounds and deep neural networks for the early detection of COVID-19," *Biomedical Engineering Advances,* vol. 3, p. 100025, 2022.

[12] R. Islam, E. Abdel-Raheem and M. Tarique, "Early Detection of COVID-19 Patients using Chromagram Features of Cough Sound Recordings with Machine Learning Algorithm," in *Proceedings of the International Conference on Microelectronics (ICM)*, New Cairo City, 2022.

[13] R. Islam, E. Abdel-Raheem and M. Tarique, "Voice pathology detection using convolutional neural networks with electroglottographic (EGG) and speech signals," *Computer Methods and Programs in Biomedicine Update,* vol. 2, p. 100074, 2022.

[14] R. Islam and M. Tarique, "A novel convolutional neural network based dysphonic voice detection using chromagram," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 12, no. 5, pp. 5511-5518, 2022.

[15] G. Kempster, B. Gerratt, K. Abbott, J. Barkmeier-Kraemer and R. Hillman, "Consensus Auditory-Perceptual Evaluation of Voice: Development of a standardized Clinical Procedure," *American Journal of Speech-Language Pathology,* vol. 18, no. 2, pp. 124-132, 2018.

[16] R. I. Titze, "Physiologic and acoustic differences between male and female voices," *Journal of Acoustic Society of America,* vol. 85, no. 4, pp. 1699-1707, 1989.

[17] M. Hirano, Y. Kakita, K. Ohmaru and S. Kurita, "Structure and Mechanical Properties of the Vocal Fold," *Speech and Language,* vol. 7, pp. 271-297, 1982.

[18] J. Kahane, "A morphological study of the human pre-pubertal and pubertal larynx," *American Journal of Anatomy,* vol. 151, no. 1, pp. 11-19, 1978.

[19] T. Giannakopouls, Introduction to Audio Analysis, Cambridge: Aacdic Press, 2014.

[20] N. Nicola, M. Fiorella, D. Spinelli and R. Fiorella, "Acoustic analysis of voice in patients treated by reconstructive subtotal laryngectomy. Evaluation and critical review," *ACTA Otorhinolaryngol Italy,* vol. 26, pp. 59-68, 2006.

[21] M. Hossain, G. Ghulam and A. Atif, "Smart healthcare monitoring: a voice pathology detection paradigm for smart cities," *Multimedia Systems,* vol. 25, pp. 565-575, 2017.

[22] R. Islam, E. Raheem and M. Tarique, "A Novel Pathological Voice Identification Technique through Simulated Cochlear Implant Processing Systems," *Applied Sciences,* vol. 12, no. 5, pp. 1-21, 2022.

[23] R. Islam, M. Tarique and E. Abdel-Raheem, "A Survey on Signal Processing Based Pathological Voice Detection Techniques," *IEEE Access,* vol. 8, pp. 66749-66776, 2020.

[24] L. Lee, L. Chamberlain, R. Loudon and J. Stemple, "Speech segment durations produced healthy and asthmatic subject," *Journal of Speech Disorder,* vol. 53, no. 2, pp. 186-193, May 1998.

[25] A. Alzheimer, "On a peculiar disease of the cerebral cortex," *Allgemeine Zietschrift fur Psychiatrie,* vol. 64, p. 146–148, 1907.

[26] J. Cummings, F. Benson, M. Hill and S. Read, "Aphasia is dementia of the Alzheimer type," *Neurology,* vol. 35, no. 3, pp. 394-397, 1985.

[27] K. Forbes, A. Venneri and M. Shanks, "Distinct patterns of spontaneous speech deterioration: an early predictor of Alzheimer's disease," *Brain and Congnition,* vol. 48, no. 2-3, pp. 356-361, 2002.

[28] V. Olga, "Dementia: Presentations, Differential Diagnosis, and Nosology," *The Johns Hopkins Series in Psychiatry and Neuroscience,* pp. 102-122, 1994.

[29] A. Kertesz, J. Appell and M. Fisman, "The dissolution of language in Alzheimer's disease," *Canadian Journal of Neurological Science,* vol. 13, pp. 415-418, 1986.

[30] K. Faber-Langendoen, J. Morris, J. Knesevich, E. LaBarge, J. Miller and L. Ber, "Aphasia in senile dementia of the Alzheimer type," *Annals of Neurology,* vol. 33, no. 4, pp. 365-370, 1988.

[31] L. Ferm, "Behavioral activities in demented geriatric patients," *Clinics,* vol. 16, no. 4, pp. 185-194, 1974.

[32] H. S. Kirshner, "Progressive aphasia and other focal presentations of Alzheimer disease, Pick disease, and other degenerative disorders," *Dementia: Presentations, differential diagnosis, and nosology,* pp. 108-122, 1994.

[33] V. Olga and B. Emery, "Language disturbance: an initial symptom of cortical degeneration and dementia," *Archives of Neurology,* vol. 41, no. 5, pp. 491-496, 1984.

[34] V. Olga and B. Emery, "Language Impairment in Dementia of the Alzheimer Type: A Hierarchical Decline?," *International Journal of Psychiatry Medicine,* vol. 30, no. 2, pp. 145-164, 2000.

[35] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *Journal of Neurology, Neurosurgery, and Psychiatry,* vol. 79, no. 4, pp. 368-376, 2008.

[36] R. A. Tahami and E. Shirva, "Voice Analysis for Detecting Parkinson's Disease using Genetic Algorithm and KNN," in *Proceedings of the 18th Iranian conference on Biomedical Engineering*, Tehran, 2011.

[37] K. Rosen, R. Kent, A. Delaney and J. Duffy, "Parametric quantitative acoustic analysis of conversation produced by speakers with dysarthria and healthy speakers," *Journal of Speech, Language, and Hearing Research,* vol. 49, no. 2, pp. 395-411, 2006.

[38] B. Harel, M. Cannizzaro and P. Snyde, "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study," *Brain and Cognition,* vol. 56, no. 1, pp. 24-29, 2004.

[39] P. A. LeWitt, "Parkinson's Disease: Etiologic Considerations," in *Parkinson's Disease and Movement Disorders. Current Clinical Practice*, New Jersy, Human Press, 2000, pp. 91-100.

[40] E. Moore, M. Clements and L. Weisse, "Investigating the Role of Glottal Features in Classifying Clinical Depression," in *Proceedings the 25th Annual International Conference of the IEEE EMBS*, Cancum, 2003.

[41] M. Alpert, E. Pouget and R. Silva, "Reflections of depression in acoustic measures of the patient's speech," *Journal of Affective Disorders,* vol. 66, no. 1, p. 59–69, 2001.

[42] A. Nilsonne, J. Sundberg, S. Ternström and A. Askenfelt, "Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression," *Journal of Acoustic Society of America,* vol. 83, no. 2, pp. 716-728, 1988.

[43] D. France, R. Shiavi, S. Silverman, M. Silverman and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Transaction on Biomedical Engineering,* vol. 47, no. 7, pp. 829-837, 2000.

[44] J. Mundt, P. Snyder, M. Cannizzaro, K. Chappie and D. Geralt, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," *Journal of Neurolinguistics,* vol. 20, no. 1, pp. 50-64, 2007.

[45] D. R. Weinberger, "Implications of normal brain development for the pathogenesis of schizophrenia," *Archives of General Psychiatry,* vol. 44, no. 7, pp. 660-669, 1987.

[46] B. Elvevåg, P. Foltz, M. Rosenstein and L. Delisi, "An automated method to analyze language use in patients with schizophrenia and their first-degree relatives," *Journal of Neurolinguistics,* vol. 23, no. 3, pp. 270-284, 2010.

[47] J. Zhang, Z. Pan, C. Gui, J. Zhu and D. Cui, "Clinical investigation of speech signal features among patients with schizophrenia," *Shanghai Archives of Psychiatry,* vol. 28, no. 2, pp. 95-102, 2016.

[48] L. Kanner, "Irrelevant and metaphorical language in early infantile autism," *The American Journal of Psychiatry,* vol. 103, no. 2, pp. 242-246, 2006.

[49] M. Hoque, K. K. L. Joseph and R. Picard, "Exploring Speech Therapy Games with Children on the Autism Spectrum," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association*, Brighton, 2009.

[50] S. E. Bryson, "Brief Report: Epidemiology of autism," *Journal of Autism and Developmental Disorder,* vol. 26, pp. 165-167, 1996.

[51] L. Shriberg, R. Paul, J. McSweeny, A. Klin, D. Cohen and F. Volkmar, "Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome," *Journal of Speech, Language, and Hearing,* vol. 44, no. 5, pp. 1097-1115, 2001.

[52] P. Boersma and D. Weenink, "PRAAT: Doing Phonetics by computer," Computer Program 2005., 2005.

[53] J. Diehl, D. Watson, L. Bennetto, J. Mcdonough and C. Gunlogson, "An acoustic analysis of prosody in high-functioning autism," *Applied Psycholinguistics,* vol. 30, no. 3, pp. 385-404, 2009.

[54] L. Anne, L. Marie-Thérèse and S. Boushaba, "Prosodic Disturbances in Autistic Children Speaking French," in *Proceedings of Speech Prosody*, Campinas, 2008.

[55] T. Oxman, S. Rosenberg, P. Schnurr and G. Tucker, "Diagnostic Classification Through Content Analysis of Patient Speech," *American Journal of Psychiatry,* vol. 145, no. 4, pp. 464-468, 1988.

[56] A. Maier, T. Haderlein, F. Stelzle, E. Nöth, F. Nkenke, A. Rosanowski, A. Schützenberger and M. Schuster, "Automatic Speech Recognition Systems for the Evaluation of Voice and Speech Disorders in Head and Neck Cancer," *EURASIP Journal on Audio, Speech, and Music Processing,* vol. 1, pp. 1-7, 2009.

[57] K. Graves, "Emotional expression and emotional recognition in breast cancer survivor," *Journal of Psychology and Health,* vol. 20, pp. 579-595, 2010.

[58] "Voice Disorders," John Hopkins Medicine, 16 June 2022. [Online]. Available: https://www.hopkinsmedicine.org/ health/ conditions-and-diseases/voice-disorders. [Accessed 2022].

[59] R. W. Schafer and A. Oppenheim, "Fourier Transform and Fourier Analysis of Signals Using the Discrete Fourier Transform," in *Discrete-Time Signal Processing*, London, Pearsons, 2009, pp. 855-859.

[60] R. Schafer and L. Rabiner, Algorithms for Estimating Speech Parameter, Upper Saddle River: Pearsons Publishers, 2011.

[61] J. Jiang and L. Tan, Subband and Wavelet based Coding, Cambridge: Academic Press, 2007.

[62] O. Buza, G. Toderean, A. Nica and A. Caruntu, "Voice Signal Processing For Speech Synthesis," in *Proceedings of the IEEE International Conference on Automation, Quality and Testing, Robotics*, Cluj-Napoca, 2006.

[63] D. 0'Shaughnessy, "Linear Predictive Coding," *IEEE Potentials,* vol. 7, no. 1, pp. 29-32, 1988.

[64] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Acoustical Society of America Journal,* vol. 88, no. 4, p. 1738–1752, 1990.

[65] N. Morgan and H. Hermansky, "RASTA Processing of Speech," *IEEE of Speech and Audio Processing,* vol. 2, no. 4, pp. 578-589, 1994.

[66] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonic-to-noise ratio of a sample sound," in *Proceedings of IFA, Institute of Phonetic Sciences, University of Amsterdam*, Amsterdam, 1993.

[67] J. P. Teixeira, D. Ferreira and S. Carneiro, "Análise acústica vocal - determinação do Jitter e Shimmer para diagnóstico de patalogias da fala," in *In 6º Congresso Luso-Moçambicano de Engenharia. Maputo*, Moçambique, 2011.

[68] H. Kasuya, S. Ogawa and Y. Kikuchi, "An adaptive comb filtering method as applied to acoustic analyses of pathological voice," in *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Tokyo, 1986.

[69] G. d. Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech and Hearing Research,* vol. 36, p. 224–266, 1993.

[70] R. Bachu, B. Adapa, S. Kopparthi, D. Buket and B. K. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in *Proceedings of the ASEE Regional Conference*, Pittsburg, 2008.

[71] M. Slaney, "Auditory Toolbox. Version 2," Interval Research Corporation, West Lafayette, 1998.

[72] V. M. Thakare and U. Shrawankar, "Techniques for feature extraction in speech recognition system: a comparative study," *International Journal of Computer Applications in Engineering, Technology and Sciences (IJCAETS),* p. 412–418, 2013.

[73] J. F. Kaiser, "Some useful properties of Teager's energy operators," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, 1993.

[74] J. F. Kaiser, "On Teager's Energy Algorithm and Its Generalization to Continuous Signals," in *Proceedings of the 4th IEEE Digital Signal Processing Workshop*, Mohonk (New Palts), 1990.

[75] S. Ben-Hur, D. H. T. Horn and a. V. Vapni, "Support vector clustering," *Journal of Machine Learning Research,* vol. 2, no. 12, p. 125–137, 2002.

[76] S.Chander and P. Vijay, "3 -Unsupervised learning methods for data clustering," in *Artificial Intelligence in Data Mining: Theory and Applications*, Cambidge, Acdemic Press, 2021, pp. 41-64.

[77] D. Reynolds, "Universal Background Model," in *Encyclopedia of Biometrics*, Boston, Springer, 2009.

[78] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE Signal Processing Magazine,* vol. 4, no. 2, pp. 4-22, 1987.

[79] L. E. Baum, T. Petrie, G. Soules and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *The Annals of Mathematical Statistics,* vol. 41, no. 1, pp. 164-171, 1970.

[80] L. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of a Markov Process," in *Proceeding of the Third Symposium on Inequalities*, Los Angeles, 1969.

[81] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview," *Neural Networks,* vol. 61, p. 85–117, 2015.

[82] Y. Bengio, "Learning Deep Architectures for AI," *Foundations and Trends in Machine Learnin,* vol. 2, no. 1, p. 1–127, 2009.

[83] Y. LeCun, "LeNet-5 convolutional neural networks," 2022. [Online]. Available: http://yann.lecun.com/exdb/lenet/. [Accessed 18 June 2022].

[84] B. A. Story and Y. Zeinali, "Competitive probabilistic neural network," *Integrated Computer Aided Engineering,* vol. 24, no. 2, pp. 105-118, 2017.

[85] G. Hinton, "Deep belief network," *Scholarpedia,* vol. 4, no. 5, p. 5947, 2009.

[86] D. F. Specht, "A general regression neural network," *IEEE Transaction on Neural Networks,* vol. 2, no. 6, pp. 568-576, 1991.

[87] N. Friedman, D. Geiger and M. Goldszmidt, "Bayesian Network Classifiers," *Machine Learning,,* vol. 29, no. 2-3, pp. 131-163, 1997.

[88] A. Freitas and C. N. Silla, "A survey of hierarchical classification across different application domains," *Data Mining and Knowledge Discovery,* vol. 22, no. 1-2, p. 31–72, 2011.

[89] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkely, 1968.

[90] J. R. Quinlan, "Introduction of decision trees," *Machine Learning,* vol. 1, no. 1, p. 81–106, 1986.

[91] G. J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, New York: Wieley Interscience, 2004.

[92] 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, p. 1–9, 2019.

[93] K. Umarani and C. M. Vikram, "Phoneme Independent Pathological Voice Detection using Wavelet Based MFCCs, GMM-SVM hybrid classifier," in *Proceedings of of the International Conference on Advances in Computing, Communications, and Informatics (ICACCI)*, Mysore, 2013.

[94] F. Amala, M. Fezari and H. Bourouba, "An Improved GMM-SVM system based on Distance Matrix for voice Pathology Detection," *An International Journal of Applied Mathematics and Information Sciences,* vol. 10, no. 3, pp. 1061-1070, 2016.

[95] A. Poorzam, M. Little, J. Jensen and M. Christensen, "A parametric approach for classification of distortions in pathological voice," in *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Calgary, 2018.

[96] D. Kim and T. J. Jun, "Pathological voice disorders classification from acoustic waveform," Korean Advanced Institite of Science and Technology, Daejeon, 2018.

[97] D. Pravena, S. Dhivya and D. Durga, "Pathological voice recognition for vocal cord diseas," *International Journal of Computer Applications,* vol. 147, no. 13, pp. 31-37, 2012.

[98] K. Umarani and C. M. Vikram, "Pathological voice analysis to detect Neurological Disorder using MFCC and SVM," *International Journal of Advanced Electrical and Electronics Engineering,* vol. 2, no. 4, pp. 87-91, 2013.

[99] V. Srinivasan, V. Ramalingam and P. Arulmozli, "Artificial Neural Network Based Pathological Voice Classification using MFCC Features," *International Journal of Science, Environment, and Technology,* vol. 3, no. 1, pp. 291-302, 2014.

[100] M. Algarbi, M. Alsulaiman, G. Muhammad, M. Zakariah, T. Mesallam, K. Malki, M. Farhat, M. Amina and B. Mohammed, "Automatic Speech Recognition of Pathological Voice," *Indian Journal of Science and Technology,* vol. 8, no. 32, pp. 1-6, 2015.

[101] M. A. Wahed, "Computer Aided Recognition of Pathological Voice," in *Proceedings of the 1st National Radio Science Conference (MRSC 2014)*, Ain Shams University, 2014.

[102] S. Costa, B. Aguiar, J. Fechine and S. Correia, "Parametric Cepstral Analysis for Pathological Voice Assessment," in *Proceedings of the SAC 2008*, Fortalex, 2008.

[103] M. Fezari, F. Amara and I. El-Emary, "Acoustic Analysis for Detection of Voice Disorder using Adaptive Features and Classifiers," in *Proceedings of the 2014 International Conference on Circuits, Systems and Control*, Interlaken, 2014.

[104] J. Wang and J. Cheolwoo, "Vocal fold disorder detection using pattern recognition," in *Proceedings of the 29th Annual International Conference of the IEEE EMBS*, Lym, 2007.

[105] L. Verde, G. D. Pietro and G. Sannino, "Voice disorder identification by using machine learning techniques," *IEEE Access,* vol. 6, pp. 16246-16255, 2018.

[106] J. Orozoco-Arrayave, E. Belalcazar-Bolanos, J. Arias-Londono, S. Skodda, J. Rusz, K. Dgray, F. Honing and E. Noth, "Characterization Methods for the Detection of Multiple Voice Disorder: Neurological, Functional, and Laryngeal Disease," *IEEE Journal of Biomedical and Health Informatics,,* vol. 19, no. 6, pp. 1820-1828, 2015.

[107] D. Michaelis, M. Frolich and H. W. Strub, "Selection and combination of acoustic features for the description of pathological voices," *Journal of Acoustical Society of America,* vol. 103, no. 3, pp. 1628-1639, 1998.

[108] S. Ali and A. Ganar, "Intelligent Pathological Voice Detection," *International Journal of Innovative Research in Technology,* vol. 5, no. 5, pp. 92-95, 2018.

[109] G. Gastellones-Dominguaz and M. Sarria-Poja, "Robust pathological voice detection based on component informatics for HMM," *NOLISP,* vol. 7015, pp. 254-261, 2011.

[110] V. Sellam and J. Jagadeesan, "Classification of Normal and Pathological Voice using SVM and RBFNN," *Journal of Signal and Information Processing,* vol. 5, pp. 1-7, 2014.

[111] M. Chopra, K. Khieu and T. Liu, "Classification and Recognition of Stuttered Speech," March 2020. [Online]. Available: http://web.stanford.edu/class/cs224s/reports/Manu_Chopra.pdf. [Accessed 18 June 2022].

[112] R. Moran, R. Reilly, P. Chazal and P. Lacy, "Telephony Based Voice Pathology Assessment Using Automated Speech Analysis," *IEEE Transaction on Biomedical Engineering,* vol. 53, no. 3, pp. 468-477, 2006.

[113] Z. Kons, A. Satt, R. Hoory, V. Uloza, E. S. Vaiciukynas, A. Gelzinis and M. Bacauskiene, "On feature extraction for voice pathology detection from speech signals," in *Proceeding of the 1st Annual Afeka-AVIOS Speech Processing Conference*, Tel Aviv, 2014.

[114] S. Bielamowicz, J. Kreiman, B. Gerratt, M. Dauer and G. Berke, "Comparison of voice analysis systems for perturbation measurement," *Journal of Speech and Hearing Research,* vol. 39, no. 1, pp. 126-134, 1996.

[115] L. Eskenazi, D. Childers and D. Hicks, "Acoustic Correlates Voice Quality," *Journal of Speech and Hearing Research,* vol. 33, no. 2, pp. 298-306, 1990.

[116] A. Sasou, "Automatic Identification of Pathological Voice Quality Based on the GRBAS Categorization," in *Proceedings of the APSIPA Annual Summit and Conference*, Kuala Lumpur, 2017.

[117] B. Sabir, F. Rouda, Y. Khazri, B. Touri and M. Moussetad, "Improved algorithm for pathological and normal voices identification," *International Journal of Electrical and Computer Engineering,* vol. 7, no. 1, pp. 238-243, 2017.

[118] S. Shinohara, Y. Omiya, M. Nakamura, N. Hagiwara, M. Higuchi, S. Mitsuyoshi and S. Tokun, "Multilingual evaluation of voice disability index using pitch rate," *Advances in Science, Technology, and Engineering Systems Journal,* vol. 2, no. 3, pp. 765-772, 2017.

[119] A. Dibazar, S. Narayan and T. Berger, "Feature analysis for automatic detection of pathological speech," in *Proceedings of the Second Joint EMBS/BMES Conference*, Houston, 2002.

[120] B. A. Gudi, H. K. Shreedar and H. Nagaraj, "Estimation of severity of speech disability through speech envelop," *An International Journal on Signal and Image Processing,* vol. 2, no. 2, pp. 26-33, 2011.

[121] A. H. Patil and V. P. Baljeker, "Classification of normal and pathological voice using TEO phase and Mel Cepstral Features," in *Proceedings of the International Conference on Signal Processing and Communications (SPCOM)*, Bengalore, 2012.

[122] M. Alsulaiman, G. Mohammed and Z. Ali, "Classifications of vocal fold disease using RASTA-PLP," in *Proceedings of the International Conference on Bioinformatics & Computational Biology*, Nevada, 2014.

[123] M. Jamaludin, S. Salleh, T. Swee, K. Ahmad, A. Ibrahim and K. Ismail, "An improved time domain pitch detection algorithm for pathological voice," *American Journal of Applied Science,* vol. 9, no. 1, pp. 93-102, 2012.

[124] B. Boyanov, B. Teston, T. Hadjitodorov, P. Mitev and D. Donsko, "A software system for pathological voice acoustic analysis," in *Proceedings of the IMEKO Measurement in Biology and Medicine*, Dubrovrnik, 1998.

[125] F. Perdigae, C. Nerves and L. Sa, "Pathological voice detection using turbulent speech segments," in *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing*, Vilamoura, 2012.

[126] H. Wu, J. Soraghan, A. Lowit and G. Di-Caterina, "A Deep Learning Method for Pathological Voice Detection Using Convolutional Deep Belief Network," in *Proceedings of the INTERSPEECH*, Hyderabad, 2018.

[127] P. Murphy, "Development of acoustic analysis techniques for use in diagnosis of vocal pathology," School of Physical Science, Dublin City University, Dublin, 2019.

[128] G. Muhammad, M. Alsulaiman, A. Mahmood and Z. Ali, "Automatic voice disorder classification using vowel formants," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, Barcelona, 2011.

[129] D. G. Jamieson and V. Parsa, "Identification of pathological voices using Glottal Noise Measures," *Journal of Speech, Language, and Hearing Research,* vol. 43, no. 2, pp. 469-485, 2000.

[130] G. Kempster, B. Gerratt, K. V. Abbott, J. Barkmeier-Kraemer and R. Hillma, "Consensus Auditory-Perceptual Evaluation of Voice: Development of a standardized Clinical Procedure," *American Journal of Speech-Language Pathology,* vol. 18, no. 2, pp. 124-132, 2018.

[131] "The Rainbow Passage: Detecting Vocal Cord Paralysis," [Online]. Available: https://www. bergerhenryent.com/the-rainbow-passage-detecting-vocal-cord-paralysis/. [Accessed 10 June 2022].

[132] R. Titze, Principles of Voice Production, Upper Saddle River: Prentice Hall, 1994.

[133] K. Wu and D. Zhang, Pathological Voice Analysis, Singapore: Springer Nature Signapore, 2020.

[134] M. Pützer and W. Barry, "Institut für Phonetik," Universität des Saarlandes, 23 June 2022. [Online]. Available: http://stimmdb.coli.uni-saarland.de/index.php4#target. [Accessed 23 June 2022].

[135] K. Elemetrics, "Voice disorder database," MIT, Boston, 1994.

[136] J. Mekyska, E. Janousova, P. Gomez-Vilda, Z. Smekal, I. Rektorova, I. Eliasova, M. Kostalova and M. Mrackova, "Robust and complex approach of pathological speech signal analysis," *Neurocomputing,* vol. 167, pp. 94-111, 2015.

[137] M. A. Little, "Center for Machine Learning and Intelligent System," UCI, 26 June 2007. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Parkinsons. [Accessed 23 June 2022].

[138] A. Tsanas and M. Little, "UCI Machine Learning Repository," University of California, 26 June 2007. [Online]. Available:

http://archive.ics.uci.edu/ml/datasets/ Parkinsons+Telemonitoring. [Accessed 23 June 2022].

[139] T. Mesallam, M. Farahat, K. Malki, M. Alsulaiman and Z. Ali, "Development of the arabic voice pathology database and its evaluation by using speech features and machine learning algorithms," *Journal of Healthcare Engineering,* vol. 2017, pp. 1-13, 2017.

[140] "Voice Disorders," ASHA, [Online]. Available: https://www.asha.org/practice-portal/clinical-topics/voice-disorders/#collapse_1. [Accessed 23 June 2022].

[141] R. H. G. Martins, "Voice Disorder: Etiology and Diagnosis," *Journal of Voice,* vol. 30, no. 6, pp. P761.E1-761.E, 2016.

[142] C. Chen, Elements of Human Voice, Singapore: World Scientific Publishing, 2016.

[143] J. Wood, T. Athanasiadis and J. Allen, "Laryngitis," *The BMJ,* vol. 349, pp. 1-6, 2014.

[144] P. Kahrilas, N. Shaheen and M. Vaezi, "American Gastroenterological Association Institute technical review on he management of gastroesophageal reflux disease," *Gastroenterology,* vol. 135, no. 4, p. 1392–1413, 2008.

[145] A. Friedman, "About Vocal Cord Polyps," Voice Surgeon Net, 23 June 2022. [Online]. Available: https://voicesurgeon.net/voice-disorders/vocal-cord-polyp/. [Accessed 23 June 2022].

[146] V. Vasconcelos, A. Gomes and C. Araújo, "Vocal Fold Polyps: Literature Review," *International Archives of Otorhinolaryngology,* vol. 23, no. 1, pp. 116-12, 2019.

[147] M. Johns, R. Sataloff, A. Merati and C. Rosen, "Shortfalls of the American Academy of Otolaryngology-Head and Neck Surgery's Clinical practice guideline:

Hoarseness (Dysphonia)," *Otolaryngology Head Neck Surgery,* vol. 143, no. 2, pp. 175-177, 2010.

[148] V. Dhillon, "Muscle Tension Dysphonia," John Hopkins University, 23 June 2022. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/muscle-tension-dysphonia. [Accessed 23 June 2022].

[149] "Vocal Cord Disorder," Harvard Health Publishing, 17 January 2022. [Online]. Available: https://www.drugs.com/health-guide/vocal-cord-disorders.html. [Accessed 23 June 2022].

[150] S. Harris and S. Caldwel, "Voice Care," The Britsih Voice Association, 23 June 2022. [Online]. Available: https://www.britishvoiceassociation.org.uk/voicecare_muscle-tension-dysphonia.htm. [Accessed 23 June 2022].

[151] T. N. Wiesel and D. Hubel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *Journal of physiology,* vol. 160, pp. 106-154, 1962.

[152] P. Kim, MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligenc, New York City: APRESS, 2017.

[153] S. Du, J. Lee, H. Li, L. Wang and X. Zhai, "Gradient Descent Finds Global Minima of Deep Neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, 2019.

[154] R. M. Rangayyan, Biomedical Signal Analysis, New York: Wiley-IEEE Press, 2001.

[155] P. Du and Y. Jiaa, "Performance measures in evaluating machine learning-based bioinformatics predictors for classifications," *Quantitative Biology,* vol. 4, no. 4, pp. 320-330, 2016.

[156] P.Alku and S. Kadiri, "Analysis and Detection of Pathological Voice Using Glottal Source Features," *IEEE Journal of Selected Topics in Signal Processing,* vol. 14, no. 2, pp. 367-379, 2020.

[157] A. Al-Nasheri, G. Muhammad, M. Alsulaiman and Z. Ali, "Investigation of Voice Pathology Detectıon and Classification on Different Frequency Regions Using Correlation Functions," *Journal of Voice,* vol. 31, no. 1, pp. 3-15, 2017.

[158] N. Narendra and P. Alku, "Glottal Source Information for Pathological Voice Detection," *IEEE Access,* vol. 8, p. 67745–67755, 2020.

[159] M. Alhussein and G. Muhammad, "Voice Pathology Detection Using Deep Learning on Mobile Healthcare Framework," *IEEE Access,* vol. 6, p. 41034–41041, 2018.

[160] S. Cosentino, T. Falk, D. McAlpine and T. Marquardt, "Cochlear Implant Filterbank Design and Optimization," *IEEE/ACM Transaction on Audio Speech Lang. Process,* vol. 22, no. 2, p. 347–353, 2014.

[161] A. Katsiamis, E. Drakakis and R. Lyo, "Practical Gammatone-Like Filters for Auditory Processing," *EUROSHIP Journal of Audio Speech Music Processing,* vol. 2007, pp. 1-15, 2007.

[162] R. Kessler and D. Schindler, "Preliminary results with the Clarion cochlear implant," *Laryngoscope,* vol. 102, no. 9, p. 1006–1013, 1992.

[163] D. K. Kessler, "The Clarion® Multi-Strategy Cochlear Implant," *Annals of Otology, Rhinology, and Laryngology,* vol. 108, pp. 8-16, 1999.

[164] R. Tyler, B. Gantz, G. Woodworth, A. Parkinson, M. Lowder and L. Schum, "Initial independent results with the Clarion cochlear implant," *Ear and Hearing,* vol. 17, no. 6, pp. 528-536, 1996.

[165] T. Bäckström, "Pre-Emphasis," Aalto University, 10 April 2019. [Online]. Available: https://wiki.aalto.fi/display/ITSP/Pre-emphasis. [Accessed 28 June 2022].

[166] P. C. Loizou, "On the number of channels needed to understand speech," *Journal of Acoustic Society of Amercia,* vol. 106, p. 2097–2103, 1999.

[167] R. Reich, *Instrument Identification Through a Simulated Cochlear Implant Processing System, Masters Thesis,* Cambridge: Massachusetts Institute of Technology, 2002.

[168] R. W. Oppenheim and A. Schafer, "Digital Filter Design Techniques," in *Digital Signal Processing*, Upper Saddle River, Prentice Hall, 1975, p. 239–250.

[169] L. H. Carney and C. T. Win, "Temporal coding of resonances by low-frequency auditory nerve fibers: Single fiber responses and a population model," *Journal of Neurophysiology,* vol. 60, p. 1653–1677, 1988.

[170] E. D. Boer and R. D. Jongh, "On cochlear encoding: Potentialities and limitations of the reverse-correlation techniques," *Journal of Acoustic Society of America,* vol. 63, no. 1, p. 115–135, 1978.

[171] J. A. Holdsworth and A. D. Patterson, "A functional model of neural activity patterns and auditory image," *Advances in Speech, Hearing and Language Processing,* vol. 3, p. 547–563, 1996.

[172] J. Holdsworth and R. Patterson, "Complex sounds and auditory image," in *Auditory Physiology and Perception*, Oxford, Pergamon, 1992, p. 429–444.

[173] M. Unoki, T. Irino, B. Glasberg, B. Moore and R. Patterson, "Comparison of the roex and gammachip filters as representations of the auditory filter," *Journal of Acoustic Society of America,* vol. 120, no. 3, p. 1474–1492, 2006.

[174] D. Schofield, "Visualizations of the Speech Based on a Model of the Peripheral Auditory System," National Physical Laboratory, Tiddington, 1985.

[175] J. Moore and R. D. Patterson, "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selecting in Hearing*, London, Academic Press, 2019, p. 123–177.

[176] A. Darling, "Properties and Implementation of the GammaTone Filter: A Tutorial," UCL Psychology and Language Science, 1991. [Online]. Available: https://www.phon.ucl.ac.uk/home/shl5/Darling1991-GammatoneFilter.pdf. [Accessed 28 June 2022].

[177] "Worldometer Coronavirus," John Hopkins University, 31 January 2020. [Online]. Available: https://www.worldometers.info/coronavirus/?utm_campaign=homeAdvegas1?. [Accessed 04 July 2022].

[178] "Novel Coronavirus Technical Guidance," WHO, 17 July 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/. [Accessed 04 July 2022].

[179] B. Udugama, P. Kadhiresan, H. Kozlowski, A. Malekjahani, M. Osborne, V. Li, H. Chen, S. Mubareka, J. Gubbay and W. Chan, "Diagnosing COVID-19: The Disease and Tools for Detection," *American Chemical Society (ACS) Nano,* vol. 14, no. 4, pp. 3822-3835, 2020.

[180] "Half of the world lacks access to essential health services," WHO, 13 December 2017. [Online]. Available: https://www.who.int/news/item/13-12-2017-world-bank-and-who-half-the-world-lacks-access-to-essential-health-services-100-million-still-pushed-into-extreme-poverty-because-of-health-expenses. [Accessed 04 July 2022].

[181] B. Chauhan, "More than virus Fear of Stigma is stopping people from getting tested," The New Indian Express, 06 August 2020. [Online]. Available:

https://www.newindianexpress.com/states/karnataka/2020/aug/06/more-than-virus-fear-of-stigma-is-stopping-people-from-getting-tested-doctors-2179656.html. [Accessed 04 July 2022].

[182] S. Kliff, "Coronavirus test cost varies widely," 16 June 2020. [Online]. Available: https://www.nytimes.com/2020/06/16/upshot/coronavirus-test-cost-varies-widely.html. [Accessed 04 July 2022].

[183] S. Yadav, M. Keerthana, D. Gope, U. Maheswari and P. Ghosh, "Analysis of acoustic features for speech sound-based classification of asthmatic and healthy subjects," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, 2020.

[184] J. Kutor, S. Balapangu, J. Adofo, A. Dellor, C. Nyakpom and B. GK., "Speech Signal Analysis as an alternative to spirometry in asthma diagnosis:," *International Journal of Speech Technology,* vol. 22, no. 3, pp. 611-620, 2019.

[185] V. Nathan, K. Vatanparvar, M. Rahman, E. Nemati and J. Kuang, "Assessment of chronic pulmonary disease patients using biomarkers from natural," in *Proceedings of the IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, Chicago, 2019.

[186] A. Imran, P. Qureshi, U. Masood, M. Riaz, K. Ali, C. John, M. Hussain and M. Nabeel, "AI4COVID: AI-enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Informatics in Medicine Unlocked,* vol. 20, pp. 1-13, 2020.

[187] D. More, "Causes of Cough," Verywellhealth, 21 April 2021. [Online]. Available: https://www.verywellhealth.com/ causes-of-cough-83024. [Accessed 04 July 2022].

[188] C. Bales, M. Nabeel, N. Charles, U. Masood, H. Qureshi, H. Farooq, I. Posokhova and I. Ali, "Can Machine Learning Be Used to Recognize and Diagnose Coughs?,"

in *Proceedings of the IEEE International Conference on E-Health and Bioengineering (EHB)*, Lasi, 2020.

[189] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. Chetupalli, R. Nirmala, P. Ghosh and S. Ganapathy, "Coswara- A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis," in *Proceedings of INTERSPEECH*, Shanghai, 2020.

[190] J. Laguarta, F. Hueto and B. Subirana, "COVID-19 Artificial Intelligence Diagnosis using Only Cough Recording," *IEEE Open Journal of Engineering in Medicine and Biology,* vol. 1, pp. 275-28, 2020.

[191] B. Subirana, C. Moreno, J. Valle, A. González, B. Vizmanos and S. Sarma, "Hi Sigma, do I have the Coronavirus?," Cornell University, 10 April 2004. [Online]. Available: https://arxiv.org/abs/2004.06510. [Accessed 04 July 2022].

[192] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta and C. Mascolo, "Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound," in *Proceedings of the ACM Knowledge Discovery and Data Mining (Health Day)*, California, 2020.

[193] 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. Schuller, "An early-stage on Intelligent Analysis of Speech under COVID19: Severity, Sleep Quality,," in *Proceedings of the INTERSPEECH*, Shanghai, 2020.

[194] B. Schuller, S. Steidl, A. Batliner, J. Krajewski, J. Epps, F. Eyben, F. Ringeval, E. Marchi and S. Schnieder, "The INTERSPEECH 2014 Computational Paralinguistic Challenge: Cognitive and Physical Load," in *Proceedings of the INTERSPEECH*, Singapore, 2014.

[195] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, J. Epps, P. Laukka, S. Narayanan and K. Truong, "The Geneva Minimalistic Acoustic

Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transaction on Affective Computing,* vol. 7, no. 2, pp. 190-202, 2015.

[196] "OpenSMILE," audEERING, 04 July 2022. [Online]. Available: https://www.audeering.com/ research/opensmile. [Accessed 04 July 2022].

[197] C. Shimon, G. Shafat, I. Dangoor and A. Ben-Shitrit, "Artificial Intelligence enabled preliminary diagnosis for COVID-19 from voice cues and questionnaires," *Journal of Acoustic Society of America,* vol. 149, no. 2, pp. 1120-1124, 2021.

[198] "LIBROSA," Audio and Music Processing in Python, 04 July 2022. [Online]. Available: https://librosa.org. [Accessed 04 July 2022].

[199] M. Asiaee, A. Vahedian-Azimi, S. Atashi, A. Keramatfar and M. Nourbakhsh, "Voice Quality Evaluation in Patients with COVID-19: An Acoustic Analysis," *Journal of Voice,* pp. 1-7, 2020.

[200] G. Pinkas, "SARS-COV-2 Detection from Voice," *IEEE Open Journal of Engineering in Medicine and Biology,* vol. 1, pp. 268-274, 2020.

[201] A. Hassan, I. Shahin and M. Alsabek, "COVID-19 Detection System Using Recurrent Neural Networks," in *Proceedings of International Conference on Communication,*, Sharjah, 2020.

[202] M. Alsabek, I. Shahin and A. Hassan, "Studying the Similarity of COVID-19 Sound based on Correlation Analysis of MFCC," in *Proceedings of International Conference on Communication, Computing, Cybersecurity, and Informatics*, Sharjah, 2020.

[203] A. Popadina, A. Salah and K. Jalal, "Voice Analysis Framework for Asthma-COVID-19 Early Diagnosis and Prediction: AI-based Mobile Cloud Computing Application," in *Proceedings of the IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, Moscow, 2021.

[204] P. Mouawad, T. Dubnov and S. Dubnov, "Robust Detection of COVID-19 in Cough Sounds Using Recurrence Dynamics and Viable Markov Model," *SN Computer Science,* vol. 2, no. 1, pp. 1-13, 2021.

[205] G. Chaudhari, X. Jiang, A. Fakhry, A. Han, J. Xiao, S. Shen and A. Khanzada, "Virufy: Global Applicability of Crowdsourced and Clinical Datasets for AI Detection of COVID-19 from Cough," Cornell University, 04 July 2022. [Online]. Available: https://arxiv.org/abs/2011.13320. [Accessed 04 July 2022].

[206] K. Asp, "AAST: The Community for Sleep Care Professionals," AAST, 20 July 2020. [Online]. Available: https://www.aastweb.org/blog/obstructive-lung-disease-vs-restrictive-lung-disease-causes-diagnosis-and-treatment-options accessed on July 06, 2022.. [Accessed 07 July 2022].

[207] F. Morris, "Spirometry in the evaluation of pulmonary function," *Western Journal of Medicine,* vol. 125, no. 2, pp. 110-118, 1978.

[208] P. Galiatsatos, "What coronavirus does to the lungs," John Hopkins University, 28 February 2022. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/what-coronavirus-does-to-the-lungs. [Accessed 6 July 2022].

[209] J. Korpas, "Analysis of the Cough Sound: an Overview," *Pulmonary Pharmacology,* vol. 9, pp. 261- 268, 1996.

[210] T. Wilson, "Virufy Database," GitHub, 18 May 2021. [Online]. Available: https://github.com/virufy/virufy-data. [Accessed 07 July 2022].

[211] T. Giannakopouls, "Audio Features," in *Introduction to Audio Analysis*, New York, Academic Press, 2014, pp. 59-98.

[212] K. Kosasih, U. Abeyratne and V. Swarnkar, "High Frequency Analysis of Cough Sounds in Pediatric Patients," in *Proceedings of the 34th Annual International Conference of the IEEE EMBS*, California, 2012.

[213] V. Despotovic, M. Ismael, M. Cornil and G. Fagherazzi, "Detection of COVID-19 from voice, cough and breathing patterns:," *Computers in Biology and Medicine,* vol. 139, pp. 1-9, 2021.

[214] J. Smith, H. Ashurst, S. Jack, A. Woodcock and J. Earis, "The description of cough sounds by healthcare professionals," *Cough,* vol. 2, no. 1, pp. 1-9, 2006.

[215] H. Merge, "COVID-19 Train Audio," GitHub, 13 October 2020. [Online]. Available: https://github.com/hernanmd/COVID-19-train-audio. [Accessed 06 July 2022].

[216] J. Baker, "Psychogenic voice disorders and traumatic stress experience: a discussion paper with two case reports," *Journal of voice,* vol. 17, no. 3, pp. 308-318, 2003.

[217] C. L. Ludlow, "Spasmodic Dysphonia: a Laryngeal Control Disorder Specific to Speech," *Journal of Neuroscience,* vol. 31, no. 3, pp. 793-397, 2011.

[218] P. Harar, J. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget and Z. Smekal, "Voice Pathology Detection using Deep Learning: A Preliminary Study," in *Proceedings of the IEEE International Conference and Workshop on Bioinspired Intelligence (IWOBI)*, Funchal, 2017.

[219] B. Gao, "Cochleagram based audio pattern separation using two-dimensional non-negative matrix factorization with automatic sparsity adaptation," *Journal of Acoustics Society America,* vol. 135, no. 3, pp. 1171-1185, 2014.

[220] R. Sharan and T. Moir, "Pseudo-color cochleagram image features and sequential feature selection," *Applied Acoustics,* vol. 140, pp. 198-204, 2018.

[221] T. Moir and R. Sharon, "Acoustic event recognition using cochleagram image and convolutional neural networks," *Applied Acoustics,* vol. 148, pp. 62-66, 2019.

[222] T. Moir and R. Sharon, "Pseudo-color cochleagram image feature and sequential feature selection for robust acoustic event recognition," *Applied Acoustics,* vol. 140, pp. 198-204, 2018.

[223] R. V. Sharon and T. J. Noir, "Cochleagram image feature for improved robustness in sound recognition," in *Proceedings of the International Conference on Digital Signal Processing*, Singapore, 2015.

[224] T. Meer and M. Buermann, "Speech recognition using very deep neural networks: Spectrogram vs Cochleagram," Tom van Meer's Lab, 2020.

[225] T. Arias-Vergara, P. Klumpp, J. Vasquez-Correa, N. Nöth, J. Orozco-Arroyave and M. Schuster, "Multichannel spectrogram for speech processing applications," *Pattern Recognition and Analysis,* vol. 24, no. 2, pp. 1-9, 2020.

[226] I. R. Titze, "Physiologic and acoustic differences between male and female voices," *Journal of Acoustic Society of America,* vol. 85, no. 4, pp. 1699-1707, 1989.

[227] M. Hirano, "The structure of the vocal fold," in *Vocal Fold Physiology*, Tokyo, University of Tokyo Press, 1983, pp. 33-43.

[228] J. Kahane, "A morphological study of the human prepubertal and pubertal larynx," *American Journal of Anatomy,* vol. 151, no. 1, pp. 11-19, 1978.

[229] W. Dou, H. Wang and R. Yang, "Cochleagram-based identification of electronically disguised voice with pitch scaling in a noisy environment," in *Proceedings of the ACM Turing Celebration Conference*, Chengdu, 2019.

[230] F. Luz and F. Haider, "Attitude recognition using multi-resolution cochleagram feature," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brighton, 2019.

[231] D. Wang and X. Zhang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Proceedings of the INTERSPEECH*, Singapore, 2014.

[232] M. Tu, X. Xie and X. Na, "Computational auditory scene analysis based voice activity detection," in *Proceedings of International Conference on Pattern Recognition*, Stockholm, 2014.

[233] R. Sharan, S. Berkovsky and S. Liu, "Voice Command Recognition using Biologically Inspired Time-Frequency Representation and Convolutional Neural Network," in *Proceedings of Annual International Conference of Engineering Biomedical Society*, Montréal, 2020.

[234] Y. Muthusamy, R. Cole and M. Slaney, "Speaker independent vowel recognition: Spectrogram vs Cocohleagram," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing,*, 1990, 1990.

[235] S. Ahmed, N. Mamun and M. Hossain, "Cochleagram Based Speaker Identification Using Noise Adapted CNN," in *Proceedings of the 5th International Conference on Electrical Engineering and Information Communication Technology*, Dhaka, 2021.

[236] G. Tamulevičius, G. Korvel, A. Yayak, P. Treigys, J. Bernatavičienė and B. Kostek, "A study of cross-linguistic speech emotion recognition based on 2D Feature Space," *Electronics,* vol. 9, no. 10, pp. 1-13, 2020.

[237] A. M. Darling, "Properties and Implementation of the GammaTone Filter: A Tutorial," UCL Psychology and Language Sciences, 31 March 1991. [Online].

Available: https://www.phon.ucl.ac.uk/home/shl5/ Darling1991-GammatoneFilter.pdf. [Accessed 29 August 2022].

[238] C. Win and L. H. Carney, "Temporal coding of resonances by low-frequency auditory nerve fibers: Single fiber responses and a population model," *Journal of Neurophysiology,* vol. 60, no. 5, pp. 1653-1677, 1988.

[239] D. Schofield, "Visualizations of the speech based on a model of the peripheral auditory system," Physics, 1985.

[240] B. Moore and R. Patterson, "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selecting in Hearing*, London, Academic Press, 1986, pp. 123-177.

[241] Y. Suzuki and H. Takeshima, "Equal-loudness level contours for pure tone under free-field listening conditions," *Journal of Acoustic Society of Japan,* vol. 116, no. 2, pp. 918-933, 2004.

[242] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics,* vol. 36, pp. 193-202, 1980.

[243] A. Zisserman and K. Simonyan, "Very deep convolutional networks for large scale image recognition," in *Proceedings of ICLR*, San Diego, 2015.

[244] A. M. Little, "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease," *IEEE Transaction on Biomedical Engineering,* vol. 564, pp. 1015-1022, 2009.

[245] Z. Dankovičová, D. Sovák, P. Drotár and L. Vokorokos, "Machine Learning Approach to Dysphonia Detection," *Applied Sciences,* vol. 8, no. 10, pp. 1-12, 2018.

[246] S. Yang, F. Zheng, X. Luo, S. Cai, Y. Wu, K. Liu, M. Wu, J. Chen and S. Krishnan, "Effective Dysphonia Detection Using Feature Dimension Reduction and Kernel

Density Estimation for Patients with Parkinson's Disease," *PLOS ONE,* vol. 9, no. 2, pp. 1-10, 2014.

[247] L. Verde, G. Pietro, M. Alrashoud, A. Ghoneim, K. Al-Mutib and G. Sannino, "Dysphonia Detection Index (DDI): A New Multi-Parametric Marker to Evaluate Voice Quality," *IEEE Access,* vol. 7, p. 55689 – 55697, 2019.

[248] J. P. Teixeiraa, "Vocal Acoustic Analysis – Classification of Dysphonic Voices with Artificial Neural Networks," in *Procedia Computer Science*, Bercelona, 2017.

APPENDICES

Appendix (A)

The Gammtone Filters and Their Properties

The impulse response of the gammatone filter is given by:

$$h(t) = ct^{n-1}e^{-2\pi bt}\cos(2\pi f_0 t + \varphi)u(t), \quad \text{(A1)}$$

Assume that the carrier component is denoted by:

$$c(t) = \cos(2\pi f_0 t + \varphi). \quad \text{(A2)}$$

Additionally, assume that the gammatone distribution function is defined by:

$$r(t) = t^{n-1}e^{-2\pi bt}u(t). \quad \text{(A3)}$$

Hence, the impulse response of the gammatone filter can be expressed as:

$$h(t) = c.\,c(t)r(t), \quad \text{(A4)}$$

We can write:

$$H(f) = c.\,[C(f) * R(f)]. \quad \text{(A5)}$$

where $C(f)$ is the Fourier transform of $c(t)$, $R(f)$ is the Fourier transform of $r(t)$, and $H(f)$ is the Fourier transform of $h(t)$.

By using the Fourier transform of known functions along with Fourier transform properties [63], we can express the Fourier transform of $r(t)$ as:

$$R(f) = \frac{(n-1)!}{(2\pi b + j2\pi f)^n}, \quad \text{(A6)}$$

Alternatively, the Fourier transform of $r(t)$ can also be expressed as:

$$R(f) = (n-1)!\,(2\pi b)^{-n}\left(1 + j\frac{f}{b}\right)^{-n}. \quad \text{(A7)}$$

Now, the Fourier transform of the carrier signal, $c(t)$, can be expressed as:

$$C(f) = \frac{1}{2}e^{j\varphi}\delta(f - f_0) + \frac{1}{2}e^{-j\varphi}\delta(f - f_0). \quad \text{(A8)}$$

Substituting the Fourier transform of the $c(t)$ and $r(t)$ in Equation (A5), we can determine the expression of $H(f)$ as:

$$H(f) = c.(n-1)!\,(2\pi b)^{-n}\left(1 + j\frac{f}{b}\right)^{-n} * \left[\frac{1}{2}e^{j\varphi}\delta(f - f_0) + \frac{1}{2}e^{-j\varphi}\delta(f + f_0)\right], \tag{A9}$$

This expression can be further simplified as:

$$H(f) = c.(n-1)!\,(2\pi b)^{-n}\left[\frac{1}{2}e^{j\varphi}\left(1 + j\frac{f - f_0}{b}\right)^{-n}\right] + c.(n-1)!\,(2\pi b)^{-n}\left[\frac{1}{2}e^{-j\varphi}\left(1 + j\frac{f + f_0}{b}\right)^{-n}\right]. \tag{A10}$$

Appendix (B)

Equivalent Rectangular Bandwidth of Gammatone Filters

By definition, the equivalent rectangular bandwidth of a filter, H_{ERB} [176], is defined as:

$$H_{ERB} = \frac{\int_{-\infty}^{+\infty} |H(f)|^2 df}{2|H(f_0)|^2}, \tag{B1}$$

where $|H(f_0)|^2$ is the maximum value of the power spectrum, which occurs at $\pm f_0$. By using the Parseval's theorem, we can write:

$$\int_{-\infty}^{+\infty} |H(f)|^2 df = \int_{-\infty}^{+\infty} |h(t)|^2 dt. \tag{B2}$$

Hence, the equivalent rectangular bandwidth can be expressed as:

$$H_{ERB} = \frac{\int_{-\infty}^{+\infty} |h(t)|^2 dt}{2|H(f_0)|^2}, \tag{B3}$$

Let us assume, $\check{h}(t) = |h(t)|^2$; hence, the equivalent rectangular bandwidth can be simplified as:

$$H_{ERB} = \frac{\int_{-\infty}^{+\infty} \check{h}(t) dt}{2|H(f_0)|^2}, \tag{B4}$$

From the definition of the Fourier transform of $\check{h}(t)$, we can write:

$$\check{H}(f) = \int_{-\infty}^{+\infty} \check{h}(t) e^{-j2\pi ft} dt, \tag{B5}$$

We can find the D.C. component of $\check{H}(f)$ by substituting $f = 0$ in Equation (B5) as:

$$\check{H}(0) = \int_{-\infty}^{+\infty} \check{h}(t) dt. \tag{B6}$$

Hence, Equation (B4) can be written as:

$$H_{ERB} = \frac{\check{H}(0)}{2|H(f_0)|^2}, \tag{B7}$$

Now, using Equation (A4), we can find the expression of $\check{h}(t)$ as:

$$\begin{aligned} \check{h}(t) &= [c.\, r(t)c(t)]^2, \\ &= c^2 r^2(t) c^2(t), \end{aligned} \tag{B8}$$

$$= c^2 \check{r}(t) \check{c}(t). \tag{B9}$$

$$\text{where } \check{r}(t) = r^2(t)$$

$$= t^{2n-2}e^{-4\pi bt}u(t). \tag{B10}$$

$$\check{c}(t) = c^2(t),$$

$$= \cos^2(2\pi f_0 t + \varphi). \tag{B11}$$

By taking Fourier transform of both sides of Equation (B9), we can write:

$$\check{H}(f) = c^2.\left[\check{R}(f) * \check{C}(f)\right]. \tag{B12}$$

where $\check{R}(f)$ is the Fourier transform of $\check{r}(t)$, and $\check{C}(f)$ is the Fourier transform of $\check{c}(t)$. Now, we need to find the Fourier transform of $\check{r}(t)$ and $\check{c}(t)$. The Fourier transform of $\check{r}(t)$ can be found as:

$$\check{R}(f) = (2n-2)!\,(4\pi b)^{-(2n-1)}\left[1 + j\frac{f}{2b}\right]^{-(2n-1)}. \tag{B13}$$

The Fourier transform of $\check{c}(t)$ can be expressed as:

$$\check{C}(f) = \left[\tfrac{1}{2}\delta(f) + \tfrac{1}{4}e^{j2\varphi}\delta(f - 2f_0) + \tfrac{1}{4}e^{-j\varphi}\delta(f + 2f_0)\right]. \tag{B14}$$

Finally, substituting $\check{R}(f)$ and $\check{C}(f)$ in Equation (B12), we find the expression of $\check{H}(f)$ as:

$$\check{H}(f) = c^2(2n-2)!\,(4\pi b)^{-(2n-1)}\left[1 + j\frac{f}{2b}\right]^{-(2n-1)} * \left[\frac{1}{2}\delta(f) + \frac{1}{4}e^{j2\varphi}\delta(f - 2f_0) + \frac{1}{4}e^{-j\varphi}\delta(f + 2f_0)\right], \tag{B15}$$

We can further simplify Equation (B15) as:

$$\check{H}(f) = c^2.(2n-2)!\,(4\pi b)^{-(2n-1)}\left\{\left[\frac{1}{2}\left(1 + j\frac{f}{2b}\right)\right]^{-(2n-1)} + \frac{1}{4}e^{j2}\left[1 + j\frac{(f - 2f_0)}{2b}\right]^{-(2n-1)} + \frac{1}{4}e^{-j2\varphi}\left[1 + j\frac{(f + 2f_0)}{2b}\right]^{-(2n-1)}\right\}, \tag{B16}$$

By substituting $f = 0$ in Equation (B16), we can find the expression of $\check{H}(0)$ as:

$$\breve{H}(0) = c^2(2n-2)!\,(4\pi b)^{-(2n-1)}\left\{\frac{1}{2} + \frac{1}{4}e^{j2\varphi}\left[1 - j\frac{f_0}{b}\right]^{-(2n-1)} + \frac{1}{4}e^{-j2\varphi}\left[1 + j\frac{f_0}{b}\right]^{-(2n-1)}\right\}. \quad \text{(B17)}$$

Now, we need to find $H(f_0)$ to substitute in Equation (B4). The expression $H(f_0)$ can be obtained by replacing f with f_0 in Equation (A10) as:

$$H(f_0) = c(n-1)!\,(2\pi b)^{-n}\left[\frac{1}{2}e^{j\varphi}\left(1 + j\frac{f_0 - f_0}{b}\right)^{-n}\right] +$$

$$c(n-1)!\,(2\pi b)^{-n}\left[e^{-j\varphi}\left(1 + j\frac{f_0+f_0}{b}\right)^{-n}\right], \quad \text{(B18)}$$

$$= c(n-1)!\,(2\pi b)^{-n}\left[\frac{1}{2}e^{j\varphi}\right] + c(n-1)!\,(2\pi b)^{-n}\left[e^{-j\varphi}\left(1 + j\frac{2f_0}{b}\right)^{-n}\right].$$

Substituting H῀(0) and |H(f_0)| in Equation (B7), we can find the final expression of the H_ERB as:

$$H_{ERB} = \frac{c^2(2n-2)!(4\pi b)^{-(2n-1)}\left\{\frac{1}{2}+\frac{1}{4}e^{j2\varphi}\left[1-j\frac{f_0}{b}\right]^{-(2n-1)}+\frac{1}{4}e^{-j2\varphi}\left[1+j\frac{f_0}{b}\right]^{-(2n-1)}\right\}}{2\left|c(n-1)!(2\pi b)^{-n}\left[\frac{1}{2}e^{j\varphi}\right]+c(n-1)!(2\pi b)^{-n}\left[e^{-j\varphi}\left(1+j\frac{2f_0}{b}\right)^{-n}\right]\right|^2}. \quad \text{(B19)}$$

The plot for the H_{ERB} with varying f_0/b is shown in Figure B1. This figure shows that the H_{ERB} varies with the f_0/b for $1 < f_0/b < 3$. However, the H_{ERB} becomes independent of f_0/b when it becomes greater than 3.

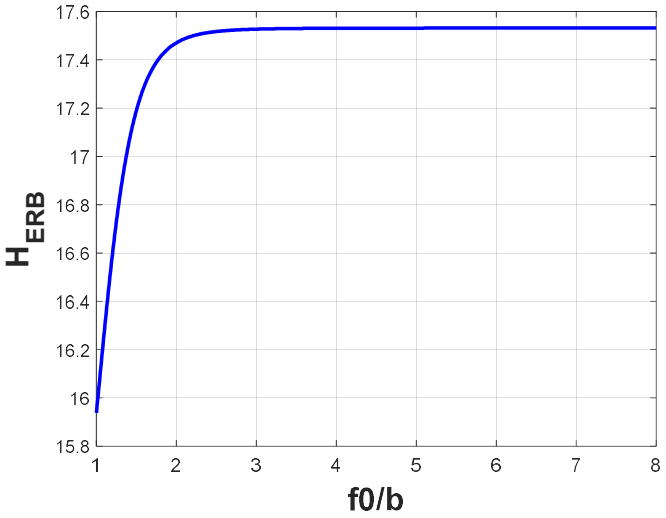


Figure B1. The variation of H_{ERB} with f_0/b. The figure shows that for $f_0/b > 3$, the H_{ERB} becomes independent of f_0/b.

VITA AUCTORIS

Rumana Islam obtained her bachelor degree (B.Sc) in Electrical and Electronics Engineering from BUET (Bangladesh University of Engineering and Technology) in 1995. She completed her Master of Science (M.Sc) in Biomedical Engineering from Wayne State University, Detroit, Michigan, USA, in 2004 on smart sensor design. She got her Doctor of Philosophy (Ph.D.) degree in Electrical and Computer Engineering, University of Windsor, Canada, in 2023. Her dissertation was on signal processing based pathological voice detection techniques. She has professional experience in both academy and industry. She served as Electrical Engineer at Prasthapana Limited, from 1995-1996. She worked as a Jr. Executive Engineer in SIEMENS Bangladesh Limited, from 1997-1998. In 1999, she joined public service as an Assistant Divisional Engineer, Planning and Development, Bangladesh Telegraph and Telephone Board (BTTB), Bangladesh. She served as an Adjunct Faculty member at East West University, Bangladesh, 2007-2008. She worked as a Lecturer in the Department of Electrical and Electronics Engineering, American International University, Bangladesh (AIUB), from 2008 to 2011. She has been serving as an Adjunct Faculty member in the Department of Electrical and Computer Engineering at the University of Science and Technology of Fujairah (USTF), UAE, since 2017. She has published several research papers in highly reputed journals and conference proceedings. Her research interests include biomedical engineering, biomedical signal processing, biomedical diagnostic tools using machine learning, digital signal processing, artificial intelligence in health informatics, biomedical image processing, wireless communication network, smart healthcare and communication system. She served as a reviewer in several referred international journals and conferences.

Published works

Journal Publications (refereed):

1. R. Islam, M. Tarique, and E. Abdel-Raheem, "A Survey on Signal Processing Based Pathological Voice Detection Techniques," IEEE Access, vol. 8, pp. 66749 – 66776, April 2020.
2. R. Islam, E. Abdel-Raheem, and M. Tarique, "Voice Pathology Detection using Convolutional Neural Networks with Electroglottographic (EGG) and Speech Signals," Computer Methods and Programs in Biomedicine Update, vol. 2, October 2022, 100074, ISSN 2666-9900, pp. 1-13, doi:. 10.1016/j.cmpbup.2022.100074
3. R. Islam, E. Abdel-Raheem, and M. Tarique, "A study of using cough sounds and deep neural networks for the early detection of COVID-19," Biomed. Eng. Adv., January 2022, vol. 3, pp. 1-12, doi: 10.1016/j.bea.2022.100025.
4. R. Islam, E. Abdel-Raheem, and M. Tarique, "A Novel Pathological Voice Identification Technique through Simulated Cochlear Implant Processing Systems," App. Scien., MDPI, vol. 12, no. 5, February 2022, pp. 1-21, doi:10.3390/app12052398.
5. Rumana Islam and Mohammed Tarique, "Chest X-ray Images to Differentiate COVID-19 from Pneumonia with Artificial Intelligence Techniques," International Journal of Biomedical Imaging, Hindawi, vol. 2022, no. 5318447, pp. 1-15, doi: https:10.1155/2022/5318447.
6. Rumana Islam and Mohammed Tarique, "A novel convolutional neural network based dysphonic voice detection algorithm using chromagram," International Journal of Electrical and Computer Engineering, vol. 12, no. 5., October 2022, pp. 5511-5518, ISSN: 2088-8708, doi: 10.11591/ijece.v12i5
7. Rumana Islam, and Mohammed Tarique,"Blind Source Separation Of Fetal ECG Using Fast Independent Component Analysis And Principle Component Analysis," INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 11, NOVEMBER 2020.

8. Mohammed tarique and Rumana Islam, "OPTIMUM NEIGHBORS FOR RESOURCE CONSTRAINED MOBILE AD HOC NETWORKS," International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) Vol.12, No.2, April 2021.
9. Mohammed Tarique and Rumana Islam, "Performances of Ad hoc Networks under Deterministic and Probabilistic Channel Conditions: Cases for Single Path and Multipath Routing Protocols", International Journal of Computer Networks and Communications, Vol. 10, No.4, July 2018, pp. 1-21
10. AKM Arifuzzaman, Rumana Islam, and Mohammed Tarique, "Window Based Smart Antenna Design for Mobile Ad hoc Networks Routing Protocol", International Journal of Wireless and Mobile Networks", Vol. 7, No. 4, August 2015, pp. 63-80
11. A.H. Siddique, AKM Arifuzzaman, Rumana Islam, and Mohammed Tarique, "Performance Study of Coded and Uncoded IEEE 802.16d under Stanford University Interim (SUI) channels", International Journal of Computer Network and Communications, Vol. 6, No.2, March 2014, pp. 143-158
12. AKM Arifuzzaman, Rumana Islam, Mohammed Tarique, and Mussab Saleh Hassan "Effects of filters on the performances DVB-T receiver", International Journal of Wireless and Mobile Networks, Vol. 5, No. 3, June 2013, pp.103-118
13. AKM Arifuzzaman, Mussab Saleh, Mohammed Tarique, and Rumana Islam, "Effects of filters on DVB-T receiver performance under AWGN, Rayleigh, and Ricean fading channels", International Journal of Computer Networks and Communication, Vo. 5, No. 4, July 2013, pp. 87-101
14. Dhrupad Debnath, Chowdhury Akram, Rumana Islam, and Mohammed Tarique, "Minimizing Shadowing effects on Mobile Ad hoc Networks", International Journal of Selected Area on Telecommunications (IJSAT), October 2011, pp. 46-51
15. Mohammed Tarique, Anwar Hossain, Rumana Islam, and C. Akram Hossain,"Issues of long-hop and short-hop routing in Mobile Ad hoc Networks: A comprehensive study", Network protocol and Algorithm, Vol. 2, No. 2, September 2010, pp. 107-131

16. Md. Anwar Hossain, Mohammed Tarique, and Rumana Islam," Shadowing Effects on Routing Protocol of Multihop Ad Hoc Networks", International Journal of Sensor, Ad Hoc, Sensor and Ubiquitous Computing, Vol. 1, No. 1, March 2010, pp. 12-28
17. Mohammed Tarique and Rumana Islam, "Minimum Energy Dynamic Source Routing Protocol for Mobile Ad hoc Networks", International Journal of Computer Science and Network Security, Vol. 7, No. 11, 2007, pp. 304-311

Conference Proceedings (refereed):

1. Rumana Islam, Esam Abdel-Raheem, and Mohammed Tarique, "Early Detection of COVID-19 Patients using Chromagram Features of Cough Sound Recordings with Machine Learning Algorithms", *The 33rd IEEE International Conference on Microelectronics, ICM 2021*, December 19-22, Cairo, Egypt, pp. 82-85.
2. Rumana Islam, Esam Abdel-Raheem, and Mohammed Tarique, "Deep Learning based Pathological Voice Detection Algorithm Using Speech and Electroglottographic (EGG) Signals," *Proceedings of the International Conference on Electrical and Computing Technologies and Applications*, November 23-25, 2022, American University of Ras Al Khaimah, United Arab Emirates (UAE).
3. Rumana Islam, Esam Abdel-Raheem, and Mohammed Tarique, "Voiced Features and Artificial Neural Network to Diagnose Parkinson's Disease Patients," *Proceedings of the International Conference on Electrical and Computing Technologies and Applications*, November 23-25, 2022, American University of Ras Al Khaimah, United Arab Emirates (UAE).
4. Rumana Islam and Mohammed Tarique, "Discriminating COVID-19 from Pneumonia Patients using Machine Learning Algorithms and Chest X-ray Images," *Proceedings of the 23rd IEEE International Conference on Industrial Technology*, August 22-25, 2022, Shanghai.
5. Rumana Islam and Mohammed Tarique, "Classifier Based Early Detection of Pathological Voice," *The IEEE International Symposium on Signal Processing and*

Information Technology (ISSPIT), Ajman University, United Arab Emirates, December 10-12, 2019.

6. Dina Taib, Mohammed Tarique, and Rumana Islam, " Voice Features Analysis for Early Detection of Voice Pathology in Children," *The IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Louisville, USA, December 6-9, 2018.
7. Tanveer Bhuiyan, Mohammed Tarique, and Rumana Islam,"Routing in Mobile Ad hoc Networks: Long hops vs short hops", *Proceedings of International Conference on Electrical and Computer Engineering*, December 2008, Dhaka, Bangladesh, pp. 600-605.