IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 22, 2020, accepted March 16, 2020, date of publication April 2, 2020, date of current version April 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2985280

A Survey on Signal Processing Based Pathological Voice Detection Techniques

RUMANA ISLAM[1], MOHAMMED TARIQUE[2], AND ESAM ABDEL-RAHEEM[1], (Senior Member, IEEE)

[1]Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON N9B 3P4, Canada
[2]Department of Electrical Engineering, University of Science and Technology of Fujairah, Fujairah 2202, UAE

Corresponding author: Rumana Islam (islamq@uwindsor.ca)

ABSTRACT Voice disability is a barrier to effective communication. Around 1.2% of the World's population is facing some form of voice disability. Surgical procedures namely laryngoscopy, laryngeal electromyography, and stroboscopy are used for voice disability diagnosis. Researchers and practitioners have been working to find alternatives to these surgical procedures. Voice sample based diagnosis is one of them. The major steps followed by these works are (a) to extract voice features from voice samples and (b) to discriminate pathological voices from normal voices by using a classifier algorithm. However, there is no consensus about the voice feature and the classifier algorithm that can provide the best accuracy in screening voice disability. Moreover, some of the works use multiple voice features and multiple classifiers to ensure high reliability. In this paper, we address these issues. The motivation of the work is to address the need for non-invasive signal processing techniques to detect voice disability in the general population. This paper conducts a survey related to voice disability detection methods. The paper contains two main parts. In the first part, we present background information including causes of voice disability, current procedures and practices, voice features, and classifiers. In the second part, we present a comprehensive survey work on voice disability detection algorithms. The issues and challenges related to the selection of voice feature and classifier algorithms have been addressed at the end of this paper.

INDEX TERMS Algorithms, issues and challenges, signal processing, surgical methods, survey, voice disability, voice features.

I. INTRODUCTION

Voice is a primitive natural tool for communication exercised by humans. Voice communication used to be an integral part of our personal and professional life. However, there are always barriers to effective voice communication. Speech impairment, due to voice disability, is one of them. An estimated 17.9 million U.S. adults of ages 18 or older report voice problems in a year [1], [2]. Among them, approximately 9.4 million adults are having problems using their voice that lasts for one week or longer [3]. According to a recent report, published by the National Center for Education Statistics, about 20% of children and youth in the age group of 3-21 years suffer from voice disability [4]. The American Speech-Language-Hearing Association suggests that voice disability occurs mainly from a disruption in the human voice generation system [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Jiafeng Xie.

The human voice generation system mainly consists of lungs, larynx, and vocal tract as shown in Fig. 1. The lungs are power sources of our voice generation system. During voice generation, we inhale air by expanding the rib cage surrounding the lungs and then we expel air from the lungs by lowering 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 a sentence or phrase. This action causes air to rush in through vocal trachea to the epiglottis. The larynx is the most complicated part of our voice generation system. It consists of cartilages, muscles, and ligaments. The main purpose of larynx is to control vocal folds, which consist of two masses stretched between the front and back of the larynx. A slit-like orifice called glottis exists between the two masses.

During normal condition, vocal folds are in a state called 'breathing'. Under a 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

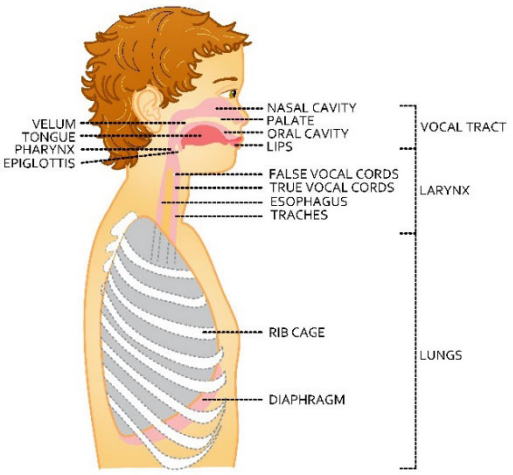


FIGURE 1. Components of the human voice generation system [5].

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 closer, become more tensed, and partially close the glottis. The partially closed glottis and increased vocal fold tension cause oscillation of the folds. The air stream from the lungs is interrupted by vocal cords and a quasi-periodic pressure wave is generated. The impulses of this pressure wave are called pitch and the frequency of the pressure 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.

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 main causes of voice disability. Some of the other causes include hearing loss, neurological disorder, brain injury, intellectual disability, drug abuse, and malfunction of the human voice system. Also, people may encounter temporary voice disorders due to allergies, large tonsils or adenoids, smoking-related illnesses, respiratory infections, and poor voice habits.

Invasive surgical procedures are commonly used to detect voice disorders. Physicians insert some kind of probe into the mouth during the endoscopic procedures namely laryngoscopy, laryngeal electromyography, and stroboscopy. These procedures are painful and often traumatize patients. Extensive researches have been conducted to find alternatives to these surgical procedures. Detecting voice pathology using voice signal processing is one of them [6].

Voice pathology detection using voice signals involve voice features extraction and analysis. Voice samples are collected in a controlled environment. Then, voice signals are analyzed to extract voice features. The next step is to classify voice samples into two categories namely normal and pathological.

While classifying the voice signals, an appropriate tool needs to be smartly selected. The most common tool is the classification algorithm. Numerous classifier algorithms have been used in the literatures. The published works show that the accuracy level highly depends on the classifier. However, there are many issues and challenges related to voice signal based pathology detection techniques. Some important issues are (a) selection of appropriate voice features, and (b) selection of an appropriate classifier tool. These issues will be discussed in this survey.

The rest of the paper is organized as follows. The medical conditions related to voice disability are discussed in section II. The current procedures and practices used to detect voice pathology are presented in section III. The voice features used in voice detection techniques have been presented in section IV. The common classifier algorithms are discussed in section V. The survey works on voice disability detection algorithms are presented in section VI. The issues and challenges related to voice pathology detection are addressed in section VII and the paper is concluded with section VIII. A list of acronyms used throughout this paper is provided in the Appendix.

II. MEDICAL CONDITIONS FOR VOICE DISABILITY

The speech pathologists have related certain medical conditions to voice disability. Some of these medical conditions include asthma, Alzheimer's disease, Parkinson's disease, depression, schizophrenia, autism, and cancer. These medical conditions are briefly explained in the following subsections.

Asthma causes swollen and inflamed vocal folds that do not vibrate properly during voice generation. This makes the voice sound hoarse and impaired. A detail investigation on this issue can be found in [7]. In the work, speech segments, of variable lengths, for asthma patients are analyzed. The speech segments include five minutes of conversation, a monologue, and counting numbers. Voice parameters namely onset time, word duration, pause time, and total activity duration for normal subjects and asthmatic subjects are considered in the work. The results show that asthmatic subjects show longer pause between speech segments, produce fewer syllables per breath, and spend a larger percentage of time in voiceless ventilator activity than their healthy counterparts.

Another major cause of voice disability is Alzheimer's disease [8]. 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 diseases also cause aphasia [9], [10]. Although memory impairment has generally been considered as the major symptom of Alzheimer's disease, it is now reported that language deficits occur in about 8%-10% of Alzheimer patients and hence they can be used as a primary symptom to detect this disease at its early stage [11]–[13]. Similar work shows that about 5% of

Alzheimer patients' language capacity steadily impairs during this disease [14]. Other works [15], [16] 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 [17].

Parkinson's disease is another major cause of voice disability. Generally, Parkinson's disease causes loss on neurons in the brain and hence affects the motor and non-motor body functions of the human body. Parkinson's disease patients face problems related to recognition, behavioral changes, insomnia, and sensory difficulties [18]. These symptoms are often followed by other symptoms including slower movement, rigidity, tremor, and postural instability. The Parkinson's disease also affects patients' muscles of voice generation system and hence patients speak slowly, loosely, and breathily. Even, they find difficulty in pronouncing words correctly. They also generate undesired voices due to their faulty vocal folds [19]–[21]. Recent research shows that voice disability can indicate an early symptom of Parkinson's disease [22].

Depression is a psychiatric disorder that affects the mood, behavior, thoughts, senses, ailments, and feelings of a human being. This disease can make a patient anxious, fatigued, irritable, and worried. The patient may have a problem in making a decision, memorizing, and losing interest in activities. Studies show that depression can also affect the patients' voice system [23]. The patients speak softly, slowly, hesitatingly, and monotonously. They often stutter and mute in the middle of a sentence [24]. Hence, voice features including pitch, energy, speaking rate, formants, and power spectral density can be used to identify a depressed patient [25], [26]. It is also shown in [27] 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 [28]. Schizophrenia patients usually suffer from delusions, hallucinations, movement disorders, and disorganized speech. They even sometimes talk about strange and unusual ideas. A study [29] shows that the fluctuations in speech can be used as biomarkers for schizophrenia. Hence, advanced signal processing techniques and artificial intelligence can be employed to investigate voice features that contain substantial emotional information of a schizophrenia patient. In [30], two spectral features namely Mel-frequency Cepstral Coefficient (MFCC) and Linear Predictive Coding (LPC) have been used to differentiate patient groups from the normal group. It is shown that MFCC scores are significantly lower, and LPC scores are significantly 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 [31]. In the work, autism is characterized by impairment in social interaction, behaviors, and communication skills. The autistic patient often says something irrelevant and it does not match with the situation [32]. Hence, speech and prosody-voice profiles can characterize the autistic patient and patients with Asperger Syndrome (AS) [33]. It is shown in the work that patients suffering from these two diseases cause residual articulation distortion errors, not understandable utterances, inappropriate in the domains of phrasing, stress, and resonance. Another study [34] correlates acoustic measurements to communication impairment due to autism. The work shows that fundamental frequency variation in the narrative of the autistic patient can be related to intelligence quotient (IQ) and verbal abilities of the autistic patient. In the work, PRAAT [35] software has been 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 compared to controlled patients. The inflection of voice, pattern of pauses, relative duration of syllables, relative loudness, and rhythm are often included in the prosodic features of voice [36]. Hence, prosody, particularly prominence and prosodic contours, can be used to investigate the communicative intent and conversational skills of autistic patients. The results presented in [37] show that abnormal prosody is the core deficit in autistic patients.

Cancer is another cause of voice disability. Study shows that voice features can be related to the cancer stage [38]. Based on the speech content analysis of 71 patients, it is shown 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 [39] to detect cancer of neck and head. The authors conclude that speech recognition can provide the percentage of correctly recognized words of a sequence. The same work shows that cancer patients have significantly lower word recognition rates than the control group. Hence, automatic speech recognition can serve as a good means to objectify and quantify cancer patients. Another study [40] shows that the role of emotional expression and progression of cancer are related. In the study, the voice samples of 25 breast cancer survivors and 25 controlled patients are used. The results show that cancer patients use significantly less inhibition words than controlled patients. The results also show that cancer diagnosis and treatment can alter the emotionally expressive behavior of a patient.

III. CURRENT MEDICAL PROCEDURES

To detect voice disability, the physicians use some common procedures namely, laryngoscopy, laryngeal electromyography, stroboscopy, and imaging tests [41]. In laryngoscopy, the throat is examined by a light source. There are three types of laryngoscopy namely direct laryngoscopy, indirect laryngoscopy, and fiber optic laryngoscopy. Direct rigid laryngoscopy procedure is used to examine the vocal cords or larynx of patients. A laryngoscope is a rigid and hollow tube with a light attached to the top. Using this tool, the physicians can examine behind the patient's tongue and down the

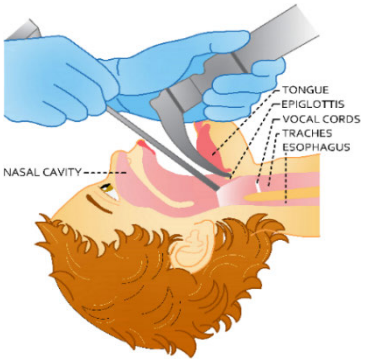


FIGURE 2. The direct laryngoscopy.

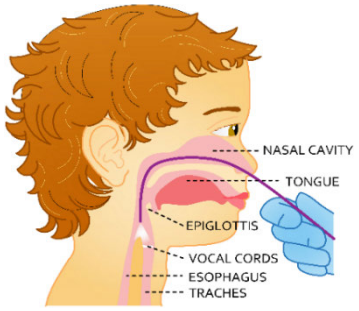


FIGURE 3. Fiber-optic laryngoscopy.

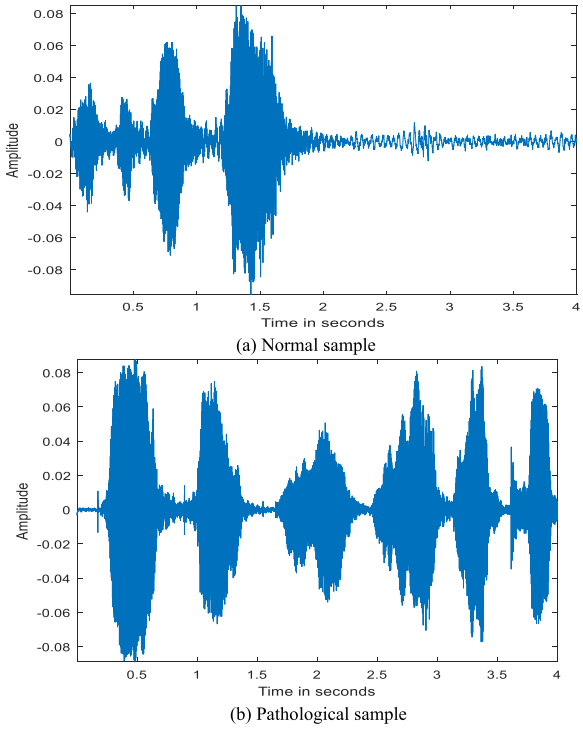


FIGURE 4. The voice samples used in the analysis.

throat to the vocal cords as shown in Fig. 2. 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. 3.

By using laryngeal electromyography (LEMG), electrical activity in the muscles of the throat is measured. LEMG is a useful diagnostic tool to examine 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 the disruption of the laryngeal nerve and superior laryngeal nerve. By using LEMG, it is possible to determine the vocal folds' tonicity. In this method, a thin needle is pierced into the neck muscles and conductivity of the muscles is measured with electrodes.

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 and this type of vibration is impossible to be noticed clearly with the naked eye. Hence, a stroboscopy is used. During this procedure, a bright flashing light is used 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. Other 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 to examine the growths of tissue or other related problems in the throat.

IV. VOICE FEATURES USED IN VOICE PATHOLOGIES

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), harmonic to noise ratio (HNR), cepstral based HNR (CHNR), noise energy to total energy ratio (NNR), zero-crossing rate (ZCR), linear frequency cepstral coefficient (LFCC), and Teager energy operator (TEO). These voice features are briefly described in the following subsection. To describe these we consider two voice samples – one for pathological baby and the other for a normal baby. These voice samples are shown in Fig. 4. 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.

A. 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 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 [5]. The voice sample $x[n]$ is first windowed with an analysis window $w[n]$ and the short-time Fourier transform (STFT), $X(n, \omega_k)$ is computed by

$$X(n, \omega_k) = \sum_{m=-\infty}^{\infty} x[m]\, w[n\text{-}m]\, e^{-j\omega_k m}, \tag{1}$$

where $\omega_k = \frac{2\pi k}{N}$ with N is the discrete Fourier transform (DFT) length. The magnitude of $X(n, \omega_k)$ 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 energies for each speech frame at time n and for l-th *mel*-scale filter is given by

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

where $V_l(\omega)$ is the frequency response of 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

$$A_l = \sum_{L_l}^{U_l} |V_l(\omega_k)|^2 \tag{3}$$

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

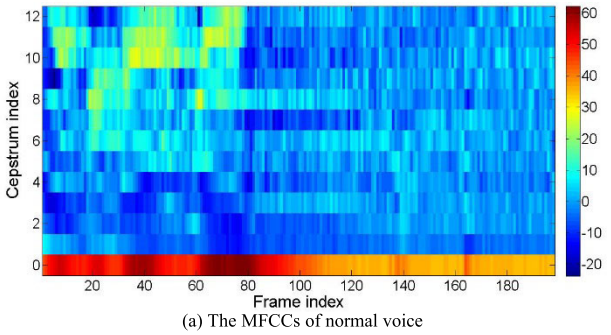
$$C_{mel[n,m]} = \frac{1}{R} \sum_{l=0}^{R-1} log\left(E_{mel}(n, l)\right) cos\frac{2\pi ml}{R}, \tag{4}$$

where R is the number of filters. An example of MFCCs of a normal voice and a pathological voice (presented in Fig. 4) are shown in Fig. 5. The plot shows the distribution of the magnitudes for MFCCs with respect to frame index and cepstrum index. It shows that the magnitude of the *mel* frequency cepstrum coefficients are high with the lower frame indices for normal voice. On the other hand, MFCCs for pathological voices are randomly distributed among a wide range of frame indices. Hence, MFCCs are extensively used in several works for discriminating pathological voice from normal voice.

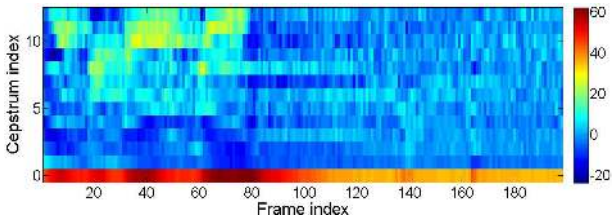
B. THE SPECTROGRAMS

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 [42]. The STFT consists of a separate Fourier transform for pieces of the waveform under a sliding window. Then, the spectrogram of the voice signal is derived from STFT by

$$S(\omega) = |X(m, \omega_k)|^2 \tag{5}$$

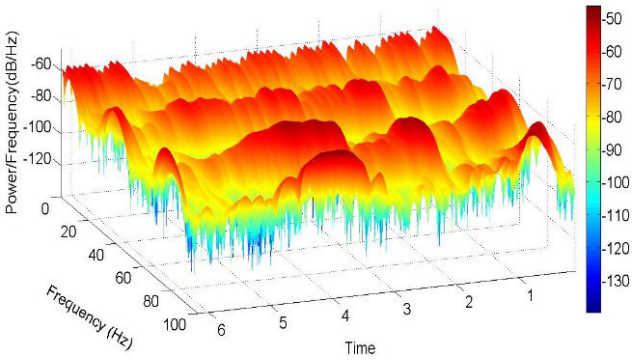


(a) The MFCCs of normal voice

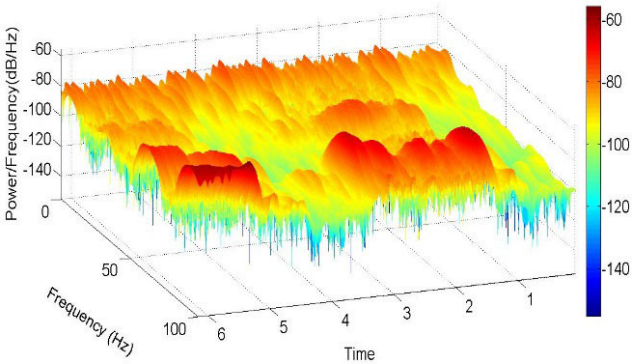


(b) The MFCCs of pathological voice

FIGURE 5. The MFCCs of normal and pathological voice samples.



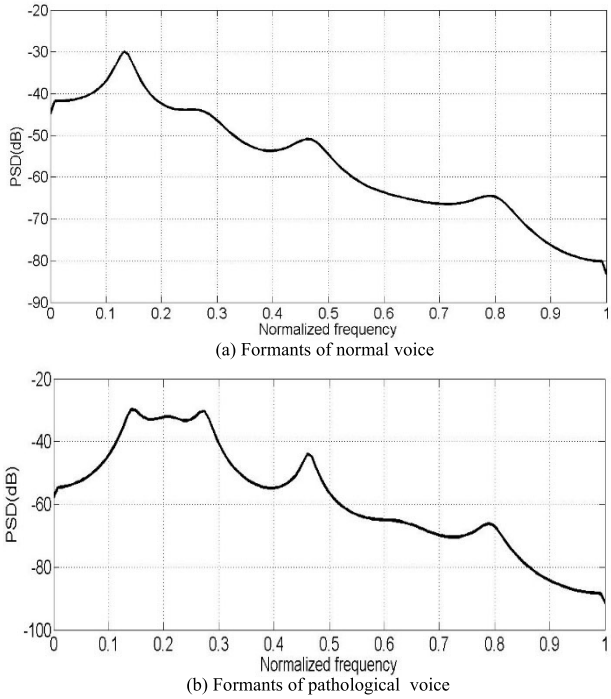
(a) Spectrogram of normal voice sample



(b) Spectrogram of pathological voice sample

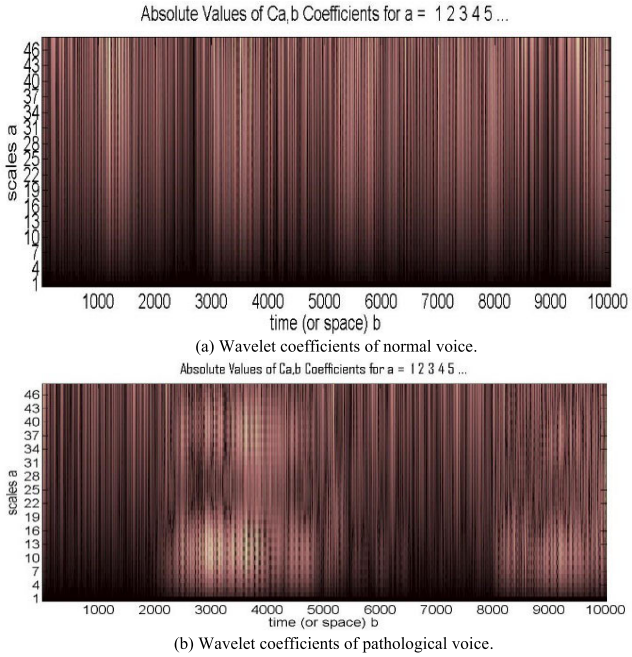
FIGURE 6. The Spectrograms of normal voice and pathological voice.

The spectrogram can be presented in 3-D plot to show the distribution of power densities with time and frequency as shown in Fig. 6. It is depicted 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 voice is uniform with respect to time and frequency. However, the same is not uniform



(a) Formants of normal voice
(b) Formants of pathological voice

FIGURE 7. The comparison of the formants.



(a) Wavelet coefficients of normal voice.
(b) Wavelet coefficients of pathological voice.

FIGURE 8. Wavelet analysis comparisons.

for pathological voice. Hence, the spectrogram is considered as a good indicator to discriminate pathological voice from normal voice.

C. FORMANTS

The formant frequency or simply formant analysis is another important voice feature investigated by the researchers. The formant frequencies are the resonance frequencies of the vocal tract and they change with different vocal tract configurations [43]. 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 frequency as shown in Fig. 7. The formant plot shows distinct peaks at certain 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 normal voice. For example, the first three peaks are closely located and are almost having the same magnitude for pathological voice. On the other hand, normal voice shows peaks that are located at almost equal distances and the peak values decrease in magnitude. Although the first formant of a normal voice carries a power similar to that of pathological voice, the other formants carry low power compared to those of pathological voice.

D. WAVELET ANALYSIS

The wavelet transform is another important tool used in voice disability detection. Its main advantage over the Fourier transform is that wavelet can provide accurate information about the fast fluctuations of signals in the time domain. It maps a time function into two functions namely scale, a and translation, b [44]. The continuous wavelet transform (CWT) of a signal $f(t)$ is defined as

$$W(a,b) = \int_{-\infty}^{\infty} f(t)\,\varphi_{ab}(t)\,d\varphi, \tag{6}$$

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

$$\varphi_{ab}(t) = \frac{1}{\sqrt{a}}\varphi\left(\frac{t-b}{a}\right) \tag{7}$$

A scaled version of the function $\varphi(t)$ with a scale factor of a is defined as $\varphi\left(\frac{t}{a}\right)$. The wavelet is a useful tool to investigate the discontinuity in pathological voice. The plot of wavelet coefficients for normal and pathological voices are shown in Fig. 8. The discontinuity in the pathological voice is more visible in the plots. Fig. 8(b) shows some discontinuity in voice signals in the range of 2500-5000 samples and 8000-8500 samples. This kind of discontinuity of voice signal does not exist in a normal voice as shown in Fig. 8(a).

E. THE LINEAR PREDICTIVE CODING (LPC)

Primarily, LPC has been introduced to compress digital signals for efficient transmission and storage. However, now LPC has become one of the most powerful speech analysis techniques and it has gained popularity as a formant estimator [45]. The LPC method is based on modeling the vocal tract as a linear all-pole infinite impulse response (IIR) filter, which is defined by

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_p(k)\,z^{-k}}, \tag{8}$$

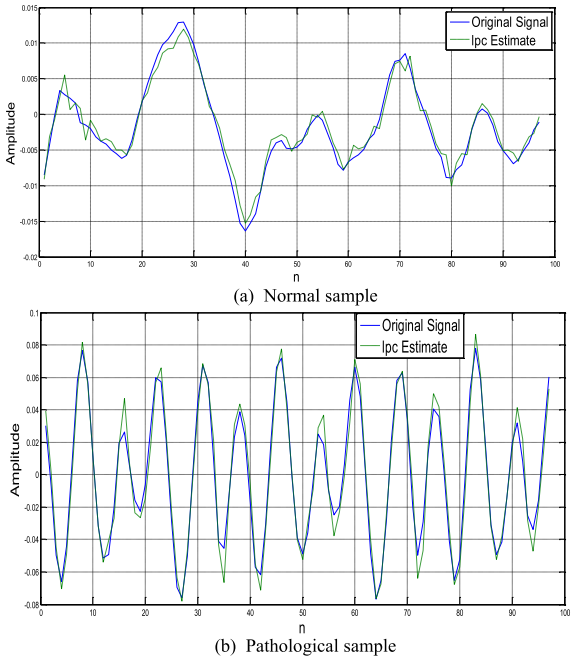


FIGURE 9. The LPC coefficients.

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 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 based on the least mean-squared error approach [46]. In this approach, the speech signal is approximated as a linear combination of its previous samples. LPC plots are generated by PRAAT [35] software and the plots (original signal and estimated signal) are shown in Fig. 9, which shows that LPC coefficients have distinctively varying magnitude in some portion of voice signal. However, the magnitude is not significant for other portion of the voice signal. The magnitude distribution can be used to differentiate pathological voice from normal voice.

F. THE PERCEPTUAL LINEAR PREDICTION (PLP)

PLP, introduced by Hermansky [47], models the human speech based on the concept of the psychophysics of hearing. The main 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 the power distribution of equal magnitude at all frequencies of the analysis band. The detailed steps of PLP computation are shown in Fig. 10.

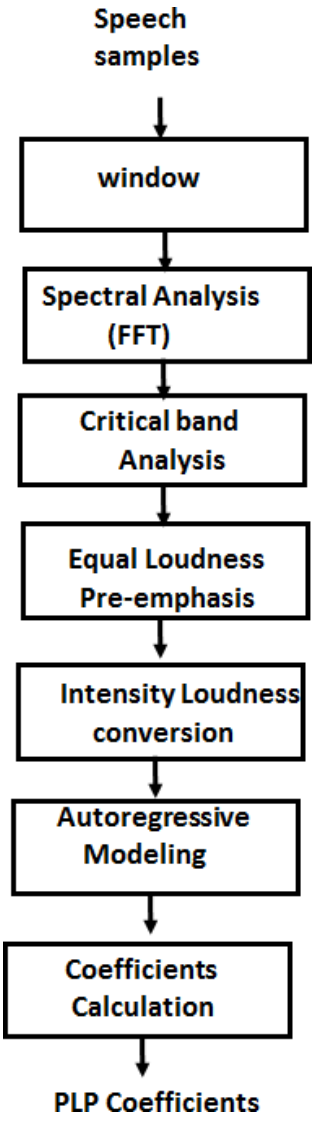


FIGURE 10. The computation of PLP [46].

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

$$P(\omega) = [Re(S(\omega))]^2 + [Im(S(\omega))]^2, \tag{9}$$

where $S(\omega)$ is the Fourier transform of the windowed voice signal. A frequency warping into the Bark scale [48] 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 [47] corresponding to an audio frequency is given by

$$R(\omega) = 6\ln\left[\frac{\omega}{1200\pi} + \sqrt{\left(\frac{\omega}{1200\pi}\right)^2 + 1}\right]. \tag{10}$$

The auditory warped spectrum is then convoluted with the power spectrum of the simulated critical-band masking curve to simulate the critical-band integration of human hearing.

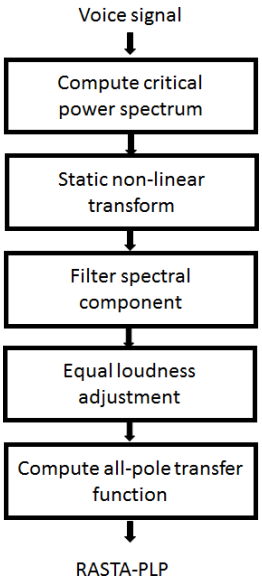


FIGURE 11. The computation of RASTA-PLP [49].

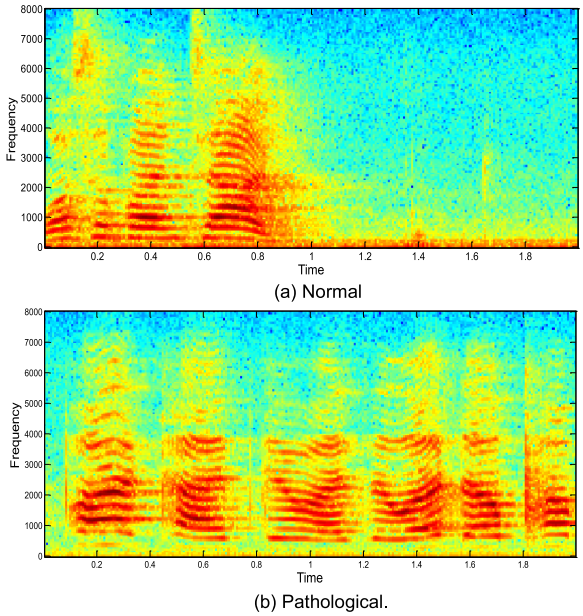


FIGURE 12. The RASTA-PLP spectra comparison.

The smoothed spectrum is then down-sampled. 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 linear predictor (LP). Finally, the cepstral coefficients are obtained from the LP coefficients by a recursive method.

G. 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 RASTA filter is used in computing the RASTA-PLP. An example of the system function for RASTA filter is defined by

$$H(z) = 0.1z^4 \cdot \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \tag{11}$$

The lower cut-off frequency of the filter determines the fastest spectral change ignored at the output. On the other hand, the high cut-off frequency determines the fastest spectral change preserved in the output. The main function of the filter is to suppress the frequency that varies more quickly or slowly in the voice signal. The steps of computing the RASTA-PLP is shown in Fig. 11.

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 voice and pathological voice are shown in Fig. 12. The RASTA-PLP plots for normal and pathological voice samples show clear differences in magnitude with respect to time and frequency.

H. 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 (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 period length, F_0 and N is the number of extracted period. 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 [50], [51].

$$Jitta = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_i - T_{i-1}| \tag{12}$$

Jitter (local): It represents the average absolute difference between two consecutive periods, divided by the average period. It is also known as *Jitt* and is given by (13), and has 1.04% as the threshold limit for detecting pathologies.

$$jitt = \frac{jitta}{\frac{1}{N} \sum_{i=1}^{N} T_i}, \tag{13}$$

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 (14) and its threshold value to detect pathologies is 0.68%.

$$rap = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_i - \left(\frac{1}{3}\sum_{n=i-1}^{i+1} T_n\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \tag{14}$$

Jitter (*ppq5*): The *ppq5* is defined by (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.

$$ppq5 = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}\left|T_i - \left(\frac{1}{5}\sum_{n=i-2}^{i+2} T_n\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \tag{15}$$

I. THE SHIMMER

The shimmer is another voice feature widely used in voice disability detection [52], [53]. Unlike Jitter, Shimmer focuses on the peak values of a signal. To determine Shimmer parameters, the algorithm begins by determining the onset time of glottal pulses of a signal and the respective magnitude of the signal at that sample. Then the algorithm is applied to determine the values of each parameter of Shimmer similarly as for Jitter. There are several Shimmer parameters as follows:

Shimmer (local): It represents 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

$$Shim = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{16}$$

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. *ShdB* (local dB) is given by

$$ShdB = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20 * log\left(\frac{A_{i+1}}{A_i}\right)\right| \tag{17}$$

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

$$apq3 = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A_i - \left(\frac{1}{3}\sum_{n=j-1}^{i+1} A_n\right)\right| \times 100}{\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{N} A_i} \tag{18}$$

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, divided by the average amplitude. The *apq5* is given by

$$apq5 = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}\left|A_i - \left(\frac{1}{5}\sum_{n=i-2}^{i+2} A_n\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{19}$$

J. NNE, GNR, HNR, AND CHNR

Normalized noise energy (NNE) is the ratio between the energy of noise and the total energy of the signal (both measured in dB) [54]. Between harmonics, the noise energy 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) voice.

The implementation of HNR is based on the mathematical fundamentals presented by Boersma [51]. It is determined by the detection of the autocorrelation function for the voice signal. The HNR is defined by

$$HNR = 10.\log\frac{AC_V\,(T)}{1 - AC_V(T)} \tag{20}$$

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 [55] 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 as 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 from turbulent noise generated in the vocal tract and is thus related to breathiness. Therefore, it is called Glottal-to-Noise Excitation Ratio (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.

K. THE ZERO CROSSING RATE (ZCR)

In the context of discrete-time signals, a zero crossing is said to occur if successive samples have different algebraic signs. The rate of zero crossings is a simple measure of the frequency content of a signal. The zero-crossing rate is a measure of the number of times in a given time interval divided by the frame that the amplitude of speech signals passes through a value of zero [56]. The zero-crossings rate is defined by

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn\left[x\left(m\right)\right] - sgn\left[x\left(m-1\right)\right]| \, w(n-m), \tag{21}$$

where $\text{sgn}\left[x\left(n\right)\right] = \left\{ \begin{array}{l} 1, \quad x\left(n\right) \geq 0 \\ -1, \; x(n) < 0 \end{array} \right\}$, and

$$w\left(n\right) = \left\{ \begin{array}{ll} \frac{1}{2N}, & 0 \leq n \leq N-1 \\ 0, & \text{otherwise} \end{array} \right\}$$

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 rate and energy distribution of a signal with respect to frequency.

L. THE LINEAR FREQUENCY CEPSTRAL COEFFICIENT (LFCCS)

LFCC is computed as MFCC with a filter bank of 40 bands MFCC-FB40 [57]. The only difference is that the *mel* frequency warping step is skipped [58]. In this algorithm, the desired frequency range is implemented 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 computation of LFCC is done in the following steps: (a) apply N-point Discrete Fourier Transform (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.

M. THE TEAGER ENERGY OPERATOR (TEO)

TEO is introduced by Teager in [59]. More works on the same topic can be found in [60], [61]. In the discrete-time domain, the Teager energy operator is given by

$$\varphi\left[x\left(n\right)\right] = x^2\left(n\right) - x\left(n+1\right)x\left(n-1\right). \tag{22}$$

There are several applications of the Teager energy operator including tracking several information sets in speech signals. This operator can track vowel and formants in the voice signal. It is also used to find the center frequency and bandwidth of the formants. A very recent application of the Teager energy operator can be found in implementing voice pathology detection algorithm.

V. THE CLASSIFIERS

An important 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 the voice disability. The classification of a given voice signal into one of many categories is very helpful in the diagnostic procedure. Pattern recognition or classification algorithms are used for this purpose. Several of classifiers have been used in voice disability detection. The most commonly used classifiers are explained in this section.

A. SUPPORT VECTOR MACHINE (SVM)

SVM applies the statistical concept of support vector to classify data [62]. 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 non-linear line. Intuitively, the performance of SVM depends on separation defined by an 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.

B. GAUSSIAN MIXTURE MODEL (GMM) AND GMM-UNIVERSAL BACKGROUND MODEL (GMM-UBM)

GMM is a probabilistic model for classifying normally distributed data within an overall data [63]. It is a widely used algorithm to classify voice features. Unlike the SVM classifier, 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 large real-world data. Particularly, 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 datasets and hence it is considered suitable in classification of large voice samples extracted from a large number of speakers [61]. 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 are estimated. In the second step, the new information regarding the parameters is mixed with old parameters and the model is updated. This kind of mixing is highly influenced by language-specific data.

C. 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 to solve such classification problems [67]. ANN possesses some properties including experience-based learning and fault tolerance. These properties make ANN particularly suitable to solve classification problems.

ANN has one hidden layer and one output layer for pattern classification. The network learns similarities among patterns directly from their instances based on an initially provided training dataset. 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 output of the network and distributed backward through the upper layers.

D. 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 [65], [66]. 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 sequence of states. Some of the common applications of HMM models include speech recognition, handwriting recognition, and gesture recognition. Recently, they are being used in voice disability detection algorithm.

E. DEEP NEURAL NETWORK (DNN)

DNN is an ANN with multiple layers [68], [69]. It can find both linear and nonlinear relationships 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 as a layer. A complex DNN uses many layers to model complex non-linear relationship between input and output. The architectures of DNN generate 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 a connection between them. The weights and inputs are multiplied and an output between '0' and '1' is returned. If the network fails to recognize a particular pattern, the algorithm adjusts weights and the whole process repeats.

F. 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 [70]. 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. CNNs have been considered very effective in computer vision. Their other applications include self-driving cars, robotics, drones, security, medical diagnosis, and treatments for visually impaired people.

G. PROBABILISTIC NEURAL NETWORK (PNN)

PNN is designed to solve classification problems using a statistical memory-based approach. It can use both supervised and unsupervised algorithm [71]. In PNN, a Parzen window is used for determining a parent probability distribution function (PDF) for each class of the population. Then, Bayes' rule is employed to allocate 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 algorithm.

With 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 a "1" and a "0" 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.

H. DEEP BELIEF NETWORK (DBN)

In machine learning, DBN is a multilayer deep neural network, with a connection between layers [72]. 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 hidden layer in subnetworks serves as a visible layer for the next layer. RBM consists 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 as the effective deep learning algorithm.

I. GENERALIZED REGRESSION NEURAL NETWORK (GRNN)

GRNN [73] 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 both linear and nonlinear regression models to predict, model, 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 pattern; 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 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.

J. BAYESIAN CLASSIFIER

The Bayesian classifier is another popular classifier used to classify data based on the common features [74], [75]. 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 the 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 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.

K. THE K-MEANS CLUSTERING

The k-means clustering is a method of vector quantization that is popularly used for cluster analysis in data mining [76]. 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-mean clustering algorithms are computationally expensive. However, some efficient heuristic algorithms have been proposed to reduce the computations by converging quickly to a local optimum. These algorithms are similar to expectation-maximization algorithm used in clustering Gaussian mixture modeling.

L. THE DECISION TREE ALGORITHM

The decision tree algorithm is a flowchart-like tree structure [77]. In decision algorithms, an internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents an outcome [77]. 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 partitions decision tree in a recursive manner. This flowchart like structure helps in decision making process similar to 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 it is faster compared to other neural networks. The complexity of decision trees lies in the number of records and the number of attributes in a given dataset. The decision tree algorithms do not depend upon probability distribution assumptions. Hence, they can handle high dimensional data with good accuracy.

M. LINEAR DISCRIMINANT ANALYSIS (LDA)

LDA is a generalization of Fisher's linear discriminant [78], which is 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 a continuous dependent variable. 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. Because they both look for linear combinations of variables that match the data. LDA is used when groups are known a priori. One of the main applications of LDA is to assess the severity state of a patient and prognosis of disease outcome. For example, an LDA classifier is commonly used in determining the severity of voice pathology into mild, moderate, and severe form.

VI. SURVEY ON VOICE PATHOLOGY DETECTION TECHNIQUES

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

A. THE MFCC TECHNIQUES

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

In [79], the authors develop a deep learning-based approach for the detection of pathological voice. In the work, normal and pathological samples of eight common clinical voice disorders are collected from a tertiary teaching

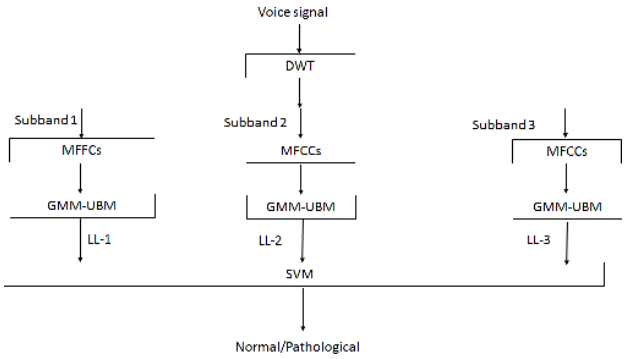


FIGURE 13. The proposed method for phoneme independent pathological voice detection [77].

hospitals. The distinct pathological voice with vocal fold nodules, polyps, cysts, neoplasm, vocal fold paralysis and adductor spasmodic dysphonia are considered in the investigation. The MFCCs are extracted from the voice samples containing sustained vowel sound for a duration of three seconds and then used in three machine learning algorithms namely DNN, SVM, and GMM using a five-fold cross-validation. To evaluate the performances of these classifiers, the authors use voice disorder database of MEEI (Massachusetts Eye and Ear Infirmary). 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 compared to other algorithms.

Wavelet sub-band based hybrid classifiers are used in [80]. Hybrid classifiers namely GMM-UBM and Gaussian Mixture Model Support Vector Machine (GMM-SVM) are used in the work. The voice samples are divided into three sub-bands using discrete wavelet transform (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. 13. The results show that the accuracy of hybrid GMM-UBM for wavelet sub band MFCCs is 96.96% that 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 of the age group 30-70 years. For each person the vowels '/a/','/i/', and '/u/', with 1.5-second duration for each, are recorded at 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 MFCCs by using GMM-UBM and are 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 1. It is shown that DB2 wavelet family provides the best accuracy

TABLE 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. 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

(i.e., 92.19%). Finally, the performance matrix for GMM-UBM and GMM-SVM are recorded in Table 2. Based on the data we can conclude 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 [81]. The authors' main focus is on the capacity of the classifier 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 advantages of generative classifiers is their capacity in handling data and in separating classes. Hence, the hybrid combination of these two types is important. The authors analyze the normal and pathological voice samples from Saarbruecken Voice Database (SVD) at the Institute of Phonetics, University of Saarland Germany. They investigate normal and pathological samples for (a) vowels with different intonations, (b) sentences (c) electroglottogram (EGG) sampled at 50 kHz at 16-bit resolution. The pathological voice sample considered in the work is neurological. Since this disease is more frequently seen among females, the authors choose a female voice database only. The major findings of the work are summarized in Table 3 and Table 4. The main focus of the work is to find a better choice of distance metric in the radial basis function (RBF) kernel. The authors have introduced two new distance metrics namely modified Kullback Leibler (KL) distance and modified Bhattacharyya distance (BH) in the paper. They have obtained an improvement of 2 % and 7 % in terms of sensitivity compared to classical KL (KL-MCS) and BH respectively.

Voice pathology due to Parkinson's disease is addressed in [82]. The proposed approach operates on cepstral features

TABLE 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 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%

extracted from voice samples using a 30 ms Hamming window. For each frame of a signal, 12 MFCCs together with log-energy are calculated. The authors argue that biomedical acoustic distortions of voice signal occur during 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 major types of acoustic distortions in the 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 voices. In the work, the authors use clean and acoustically distorted pathological voices and they achieve an 88% overall classifier accuracy.

To diagnose vocal pathology, a computerized classification model is presented in [83]. The authors use the state of the art machine learning algorithms and various classifiers in the 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%, Recurrent Neural Network (RNN) achieves an accuracy of 30% and a random forecast approach achieves an accuracy of 68%. Based on the results, the authors conclude that frequency domain voice features are more appropriate to detect voice pathology compared to its time-domain counterparts.

In [84], the authors claim that most of the vocal fold pathologies cause changes in the voice signal. Therefore, voice signals can be a useful tool to diagnose them. The paper presents a vocal fold pathology detection technique with the aid of voice signal processing. The authors first extract MFCC voice features, then they classify the feature vector using GMM. The authors also present the design and implementation of their system in the work. They show that their proposed method is less computationally complex compared to other related algorithms. The experiment is conducted on 30 speakers and the speech duration is 60 seconds. The signal processing steps performed in the work is shown in Fig. 14.

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

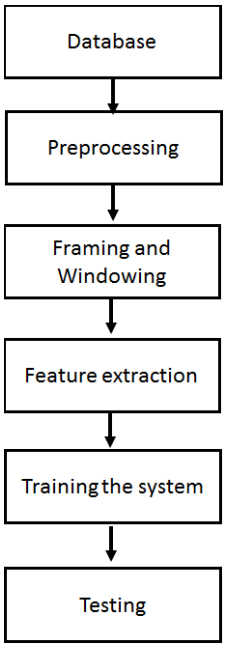


FIGURE 14. The signal processing steps used in [84].

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, a decision is made regarding normal and pathological samples. The performance matrix shows that GMM of the highest order gives the best accuracy.

In [52], the authors argue that problems like a brain tumor, lesions, neural degeneration, and brain injury may 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 the pathological condition of the brain. The work investigates the adaptation of MFCC and SVM for the diagnosis. The voice signal of 1.5-second duration is segmented by a Hamming window of 20 ms with overlays of 10 ms. Thus, 149 frames are generated and for each frame 13 MFCCs are computed. The authors use normal and pathological subjects with multiple voice disorders to test and train the SVM classifier. The accuracy level is significantly high with SVM.

A method for identification and classification of pathological voice using ANN is presented in [53]. Several other classifier algorithms namely Multilayer perceptron neural network (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 PNN behaves in a similar way to GRNN. It is also found that MLPNN performs better than PNN and GRNN in the classification of pathological voices using MFCC features.

MFCC based voice disability detection algorithms are summarized in Table 5. Based on the data listed in Table 5, we can conclude the followings. Most of the works use 'vowels' for generating a voice sample. SVM is the most

TABLE 5. Summary of MFCC based techniques.

Research works	Samples	Phonemes	Pathological Condition	Classifier	Summary of findings
Shih-Hau Fang [79]	Normal: 60 Pathological: 402	Vowels	Structural lesions, neoplasm, neurological disorder	SVM, GMM, DNN	- SVM outperforms GMM. - DNN provides the highest accuracy
Vikram and Umarani [80]	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 performances.
Fethi. Mohammed, and Hocine [81]	Samples: 200	Vowels '/a/', '/i/', and '/u/', " Good morning", "How are you?"	Spasmodic Dysphonia	GMM-HMM , SVM-ANN, GMM-SVM	- GMM-SVM outperforms other two algorithms provided distance metric is suitably chosen.
Amir Hussain *et al.* [82]	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 SVM classifier can be improved by 11% if noise information is used.
Tae Joon [83]	Not mentioned	General voice samples	Neoplasm, phono-trauma, vocal palsy	DNRNN RNN	- Highest accuracy is achieved with DNRNN
Paravena *et al.* [84]	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 [52]	Normal: 56 Pathological: 67	Vowel '/ah/'	Parkinson's disease, Vocal cord paralysis, cerebral demyelination	SVM	- Highest accuracy of 93%
V. Srinivasan [53]	Samples: 20	General voice sample	General	Multilayer Perceptron Neural Networks, Generalized regression Neural Networks	- MLPNN achieves an accuracy of 100%

popular algorithm used in MFCC based voice disability detection algorithms and MLPNN classifier achieves the best accuracy (i.e., an accuracy of 100%).

B. THE MULTIPLE FEATURES

The main motivation for using multiple voice features is to improve detection accuracy. The researches 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 Hidden Markov Model Tool Kit (HTK) is used in [85] for identifying pathological voice. By using HTK, the highest accuracy achieved is 94.44% for normal voice and 88.63% for pathological voice. The authors develop their algorithm based on the HTK tool. In their algorithm, voice features including MFCCs, PLP, RASTA-PLP and LPC are used. In their analysis, they consider voice samples of 297 speakers — 121 of them are normal and the remaining 176 have five types of vocal fold disorders. The results show that the best accuracy achieved 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 5-fold cross-validation. 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 [86], the authors present a computer-based algorithm for classifying a pathological voice from a normal voice. In the work, 50 voice samples are investigated (20 normal samples and 30 pathological samples). The features used in

TABLE 6. Accuracy, sensitivity, and specificity as a summary of single features and combined ones.

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 mean and energy mean		85%	87%	86%
Energy mean and MFCC		55%	90%	76%

the work 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 normal and pathological voices. The work is focused on detecting laryngeal voice disorder. The results of the work are summarized in Table 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.

LPC based cepstral analysis is used to discriminate pathological voices in [87]. The main focus is to detect vocal fold edema. The voice features used in the work include cepstral (CEP), delta cepstral (DCE), weighted cepstral (WCEP) and weighted delta cepstral (WDCEP) coefficients. 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 the works are 53 patients (21 male, and 32 female). In this work, all normal voice samples are down sampled to 25 kHz. The database contain more than 1400 voice samples with sustained vowel '/a/' from around 700 subjects. The results of the work are summarized in Table 7. The authors also present the ROC curve for all coefficients. The results show that DCE provides the best efficiency in terms of pathological voice detection; however, CEP provides the best correct acceptance rate.

A supervised algorithm is used in [88] to classify pathological voice from a normal voice. The voice features that are considered in the work are MFCCs and energy variation of Jitter and Shimmer. The authors classify the data using GMM. The procedure used in the work is illustrated

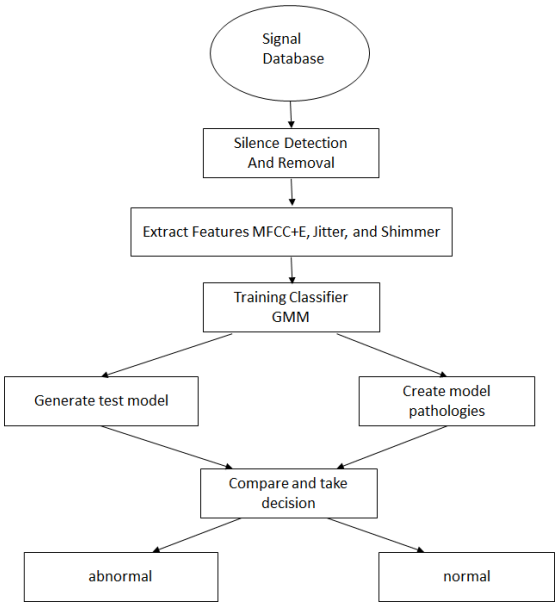


FIGURE 15. Pathological voice detection by GMM [88].

in Fig. 15. The results presented in [88] are summarized in Table 8 and Table 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 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 [89]. 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 a normal voice from a pathological voice. The authors discover that GMM based method can provide superior classifications rate compared to other classification methods. In the study, the authors consider cases with vocal fold diseases namely Cyst, Edema, Laryngitis, Nodule, Palsy, Polyp, and Glottis cancer.

Detection of dysphonia is addressed in [90]. The 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 by mainly 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, noise-to-harmonic ratio (NHR), SPI, APQ, RAP, spectral features, perturbation, noise, and energy parameters are used. Several machine learning algorithms including SVM, Decision Tree, Bayesian classification, logistic model tree, and instance-based learning are used in the work. The results are compared in terms

TABLE 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

TABLE 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 9. The confusions 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%

of accuracy, sensitivity, specificity, and receiver operating characteristic. The results show that the best accuracy is achieved by the SVM or decision tree algorithm.

The work, presented in [91], evaluates the accuracy of different characterization methods for automatic detection of multiple speech disorders. The pathologies considered in the paper include dysphonia due to Parkinson's disease, laryngeal pathologies, and hyper nasality in children with cleft lip and palate. The authors use four different methods namely noise content measure, spectral-cepstral modeling, non-linear features, and stability in fundamental frequency. The authors conclude that stability measure is suitable for Parkinson's disease and laryngeal pathologies. The spectral cepstral features are suitable for the detection of hyper nasal voice. Noise measures are suitable for dysphonic voices. The authors also conclude that individual feature is not suitable equally to model all voice pathologies. Hence, it is important to study the physiology of each impairment to choose the most appropriate set of features.

In [92], Jitter, Shimmer, periodic correlation, and GNE are used for the detection of voice pathology. An additional feature namely the noise content of a speech signal is used in the work. The authors argue that GNE is an acoustic measure that has advantages over NNE or Cepstrum based harmonics to noise ratio. Because GNE is found to be independent of variations of the fundamental frequency (Jitter) 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 in differentiating various phonation mechanisms and specific vocal pathologies as well as in monitoring progress of voices during voice rehabilitation.

Detection of vocal fold pathology with the aid of speech signal recorded from the patients is presented in [93]. 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 aim of the work presented in [94] is to compare and evaluate dynamic feature sets that are suitable for the classification of pathological voice using HMM. The features used to model speech are MFCC, HNR, GNE, NNE, and energy envelopes. The feature extraction is carried out using 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 using PCA reduces the dimension of 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 [95], explores and compares various classification models to find the ability of acoustic parameters in differentiating normal voices from pathological voices. The authors use different classification algorithm namely SVM and Radial Basis Functional Neural Networks (RBFNN). Acoustic parameters used in the work include signal energy, pitch, formant frequencies, mean square residual signal, reflection coefficients, Jitter and Shimmer. In the study, various acoustic features are combined to form a feature set. The results show that an accuracy of 91% can be achieved by the RBFNN algorithm compared to an accuracy of 83% that can be achieved by SVM.

Stuttering voice disability is addressed in [96]. The authors claim that over 3 million American stutter when they speak and many voice interfaces that exist with the consumer technology often neglect 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%. In the work, 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 as well as study techniques on applying these classifiers to automatic speech recognition mode. It is shown that the classifier can effectively parse out stuttered speech before processing the same. The classification algorithm used in the work 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 [97]. The system uses a linear classifier to process 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. In the work, the authors classify voices into four subclasses namely 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 [98], the authors argue that pathological voice detection algorithms often fail to correctly detect voice pathology. Additionally, classification rates are still insufficient for reliable and large scale screening. The work reviews performance of state-of-the-art methods and their weaknesses. The authors include the features in the time and frequency domain. The features are evaluated by different machine learning techniques. Based on the results, they conclude that the spectral features are the most important features. On the other hand, pitch related features are least 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 [99], the authors investigate dysphonic voices. Sustained vowels from male and female speakers with mild to severe dysphonia are analyzed in the work. Multiple voice features are used in the work. The authors make several important conclusions in the paper. 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.

Vocal disorders are investigated in [100]. The work investigates five-voice qualities. Six acoustic measures are examined in the work. The authors extract all the measures from residue signals obtained by inverse filtering the speech signal using the LPC technique. The authors conclude that pitch amplitude (PA) and HNR are the two most useful parameters for predicting vocal quality.

TABLE 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 main focus of the work presented in [101] is on the automatic assessment of pathological voice quality by identifying four attributes based on Grade Roughness Breathiness Strain (GRBS) categorization. The proposed method adopts higher-order local autocorrelation (HLAC) features, which 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 [102], the authors argue that although there are many research works published to detect pathological voices; however, only a few of them deal with the severity of estimation of voice disabilities. The authors present an automatic classifier using an acoustical measurement of sustained vowel '/a/' and pattern recognition tool based on neural networks. In the work, the authors include four acoustic features. The degree of severity is estimated depending on how these parameters are far from standard values. In the analysis, the authors use healthy and pathological voice samples from the German database. 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 pathological class. The authors recommend the values, shown in Table 10, to differentiate between pathological and normal voices.

The research, presented in [103], compares the effectiveness of pitch rate, Jitter, Shimmer, and harmonics-to-noise ratio as indices of voice disability in English, German, and Japanese language speakers. This study includes recitation of a page instead of using only long vowel sounds. The results show that for English, Jitter, Shimmer, and HNR are effective indices for long 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 illness affecting their voices. The reading speech results in German, Japanese, and English are similar. The pitch rate shows the greatest efficiency for identification.

An automatic speech recognition system using HTK is presented in [85]. The authors suggest that the voice produced by a pathological patient is not like a normal speaker due to irregular vibration and incomplete closure of vocal fold. Four voice features are used in the work. The voice samples of 297 speakers are used. Among them 121 are normal

TABLE 11. The classifications using MDVP.

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

TABLE 12. The classifications using HMM with MFCC and MFCC+Pitch.

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

speakers and the rest of them are pathological patients with vocal fold disorders. The authors have used a fuzzy controller for automatic segmentation of normal and pathological voice. The authors also suggest that a genetic algorithm and other optimization techniques can be used to improve the performance of the fuzzy logic control algorithm.

In a study, presented in [104], the authors present a robust, rapid, and accurate system for automatic detection of normal and pathological speech. The system uses fully automated measures of vocal tract characteristics and excitation information. In the work, the authors use MFCC coefficients and pitch dynamics to model the Gaussian mixture in the HMM classifier. The authors compare their work with the existing best performing work and it is shown that their method outperforms other classifiers by 8%. In the paper, the authors use two methods namely Multi-dimensional Voice Program (MDVP) and HMM. The results, summarized in Table 11 and 12, show that GMM provides the highest accuracy using MDVP. However, the accuracy is 99.44% when MFCC and Pitch are combined.

The summary of the mixed features-based voice disability detection algorithms is presented in Table 13. Based on the table, we can conclude the following. Although more than one voice features have been used, MFCC is one of the most common features used and most of the works use multiple classifiers. Mostly vowels are used for generating voice samples. Among the classifiers, SVM and ANN are commonly used in multiple features based algorithms.

C. TIME DOMAIN

In voice disability detection algorithms, voice features, other than 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.

The main focus of the work presented in [105] 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 in the age group of 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 classify the voice disability into three levels.

It is claimed in [106] 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 performance. The paper presents an approach that improves the standard scheme of HMM-based classifier to detect voice pathologies. The authors use HMMs to derive discriminative voice features defined by specific components. The results show that the proposed system significantly outperforms other classification systems. It achieves high accuracy using a relatively simpler procedure to generate an optimal decision boundary.

In [107], the authors use a new feature called the TEO phase for automatic detection of normal and pathological voices. The authors get the idea of TEO from a work that used the LP residual phase for speaker recognition. The authors use second-order polynomial classifier on a subject. They also use two different methods namely the TEO phase and score level fusions in the work. The comparison of the two methods is listed in Table 14 in terms of classification accuracy (ACC) and equal error rate (EER).

RASTA-PLP is used in [108] to identify four different types of vocal fold disorders. In this study, dysphonic patients consisting of 40 males and 20 females are investigated. The diseases are classified by using a multi-class SVM. The results show that a 100% classification rate can be achieved by choosing a suitable word for each disease. In the work, RASTA-PLP voice features are first extracted from the voice samples. Then the voice features are compressed by using a vector quantization, which is implemented by using the k-mean algorithm.

Table 15 presents the summary of time-domain features-based voice disability detection algorithms. Based on the data presented in the table, we can conclude the followings. The highest accuracy obtained is 100% for a specific pathology. The vowels and other native words have been used for generating voice samples.

D. THE PITCH

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

In [109], the authors present a new pitch detection algorithm that is suitable for detecting pathological voices. The main target of the work is to detect dysphonia. The proposed method uses the frame size of half-way 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 method is more reliable

TABLE 13. Summary of mixed features based classification.

Research works	Samples	Phonemes	Pathological Condition	Features	Tools	Summary of findings
Manu chopra[96]	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 [97]	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 of 87% for neuromuscular, 78% abnormality and 61% mixed pathology
Zvi Kons [98]	Samples: 320 males and 339 females	Vowel '/a/' 2-5 second	Nodule, Polyp, Cyst, Cancer, Polypoid, Hyperphasia, 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.
Algubric [85]	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 [99]	Male: 29 Female: 21	Vowel '/a/'	Dysphonia	Fundamental frequency, Shimmer, Jitter, harmonics, signal to noise ratio, and HNR	CSpeech, Computerized speech laboratory (CSL), and Soundscope	- Wide variation of reliability with increased pathology using Shimmer.
L. Eskenazi [100]	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 [101]	Pathological: 60	Japanese Vowel	Roughness, Breathiness, Asthma, and Strain	Higher-Order Local Autocorrelation (HLAC)	FFNN, AR-HMM	- An accuracy of 87.75% is achieved to detect voice pathology.
Brahim *et al.* [102]	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 [103]	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 [104]	700 samples MEEI	Vowel '/a/' Rainbow passage	Organic, Neurologic, Traumatic, and Psychogenic	MFCC, Pitch	GMM, MDVP	- An accuracy of 99.44% is achieved.

TABLE 14. The comparison between 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

to detect pitch, either 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 dictates 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 [110]. The software system can evaluate pitch period, degree of unvoiceness (DUV), pitch perturbation quotients (PPQ), and 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 (CQ). The results show that the software can detect pathological voices by examining the above-mentioned voice features.

Unlike other works, unvoiced part of voice samples is investigated in [111]. The authors argue that most of the existing works depend on the voiced part of a speech sample to detect voice pathology and these works use a pitch detector to separate 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/unvoiced part. The results show that useful pathological information is indeed present in turbulent or near unvoiced segments.

The summary of the pitch based voice disability detection algorithms is presented in Table 16. Based on the data, we can conclude the followings. The pitch feature is very useful to detect voice disabilities. Although PRAAT is widely used in voice pathology detection; however, some algorithms in time domain can provide even better accuracy for dysphonia, laryngeal, and neurological voice disorder. The vowels are mostly used in the analysis.

E. THE SPECTROGRAM FEATURES

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

Pathological voice disorder, due to vocal cord paralysis or Reinke's edema, is investigated in [112]. In the paper,

TABLE 15. Summary of time domain features.

Research works	Samples	Phonemes	Pathological Condition	Features	Tools	Summary of findings
Anandthirtha [105]	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[106]	Pathology: 65 Normal: 13	Spanish vowel	Dysphonia, Hypernasality, and Dysarthria	Nonlinear parameter, entropy	HMM tools	- Accuracy of 99% is obtained.
Hemant [107]	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 [108]	65 samples	Vowel '/a/'and '/i/'	Cyst, GERD, Polyp, and Sulcus	PLP, RAST-PLP	SVM	- 100% accuracy to classify GERD and polyp. - Maximum accuracy of 75% and 83% for cyst and sulcus respectively.

TABLE 16. Summary of pitch based voice disability detection algorithms.

Research works	Samples	Phonemes	Pathological Condition	Features	Tools	Summary of findings
Mohammad Redzuan *et al.* [109]	Normal: 49 Pathological: 87	Vowel '/a/'	Dysphonia	Pitch	Pitch Detection Algorithm (PDA)	- The proposed algorithm performs better than PRAAT in detecting voice pathology..
Boynov [110]	Normal: 100 Pathological: 300	Vowel '/a/'	Laryngeal, Neurological	Pitch	ANOVA	- Very significant changes in DH, DPP, DUV, APR and PECM
Fernando [111]	Normal: 53 Pathological: 660	Vowel '/a/' and Rainbow passage	General	Pitch	Multilayer Perceptron	- The highly turbulent speech contain useful pathological information.

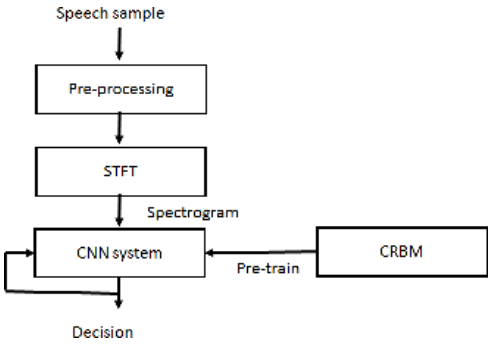


FIGURE 16. The signal processing steps used in [112].

the authors claim that 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 supervised back propagation learning algorithm to do fine tuning of the weights. The signal processing steps are shown in Fig. 16.

In [113], the authors argue that the most commonly 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 causes increased 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 the 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 17. We can conclude the following based on the data. Vowels are mostly used as voice samples. A deep learning algorithm is helpful for detecting voice disability. Jitter and Shimmer adversely affect the voice disability decision.

F. 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.

The first two formants of vowels are used in [114] for voice disorder classification. Five voice disabilities are addressed in the work. The four features are used by two types of classifier namely vector quantization and neural network. The results show that neural networks perform better than vector quantization in terms of accuracy.

Four fundamental frequencies (F0) and two F0- independent measures are used to quantify pathological voice [115]. Two of F0-dependent measures are computed in the time domain, and two others are computed using spectral information from a vowel. The F0-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 F0-independent. This eliminates the need for a computationally intensive high precision F0 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 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.

VII. ISSUES AND CHALLENGES OF VOICE DISABILITY DETECTION ALGORITHMS

Voice disability detection is usually initiated by using a screening method after receiving concern from patients, parents, teachers, and healthcare service providers. During the

TABLE 17. Summary of spectrogram based voice disability detection algorithms.

Research works	Samples	Phonemes	Pathological Condition	Features	Tools	Summary of findings
Huiyi [112]	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 algorithm can be trained with a small amount of data.
Peter Murphy [113]	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.

TABLE 18. Summary of formants based voice disability detection algorithms.

Research works	Samples	Phonemes	Pathological Condition	Features	Tools	Summary of findings
Vijay Persa [115]	Normal: 53 Pathological: 175	Vowels	Vocal noise	Formants	LP	- The vocal noise is related to LP model output.
Gulam Mohammad [114]	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%

screening, any deviation from a 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. Also, other voice features including MFCC, spectrogram, formants, wavelets, LPC, PLP, RASTA-PLP, Jitter, Shimmer, GNR, HNR, CHNR, NNR, ZCR, LFCC, and Teager energies are popularly used in voice pathology detection. Primarily, from voice sample collection and assessment to final classification stage, the following issues need to be considered.

A. SAMPLE COLLECTION ENVIRONMENT

Voice samples must be collected and assessed in a controlled environment [116]. It is suggested that voice data must be collected in a quiet environment. The other requirements are: (i) 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.

B. VOICE SAMPLES

There is no consensus about the most representable voice samples. However, most of the voice detection algorithms use common vowels. The rest of the works use sentences and running speeches. The followings are recommended in [116].

- *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 a tense vowel. It is also recommended that the patients should be asked to say vowel '/a/' for a sustained period of 3-5 second. Then the patient should be asked to say 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 that have been recommended in [112] are: (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 nasal sound, and the sixth sentence is mostly voiceless. In addition to these sentences, other works have used the ''Rainbow Passage'' [117] for voice disability detection. The specialists use this passage to

diagnose a patient, who is suffering from vocal cord paralysis or vocal cord paresis. This passage is considered suitable to assess the mobility of vocal cords for a patient.

- *Running speech* The clinicians urges 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.

C. THE DATA SOURCES AND SAMPLE

The common sources of voice samples are the local clinics. One of the main sources of a database is Massachusetts Eye and Ear Infirmary (MEEI) Voice Disorders Database [118]. However, voice recording environment and voice recording techniques are not mentioned in the database. Hence, these are also important aspects that need to be considered in implementing voice pathology detection algorithms.

D. SAMPLE SIZE

The data sample size is 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 very large samples. For example, only a few voice samples (i.e., 20) are used in [53]. On the other hand, large samples (i.e., 3750) are analyzed in [82]. Although it is recommended to use large samples for training and classification, there are no general recommendations about the sample size.

E. 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 [105]–[108].

F. CLASSIFICATION ALGORITHMS

Several classification algorithms have been used by the 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. Also, 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.

G. 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 elicit 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.

H. VOICE PATHOLOGY

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

VIII. CONCLUSION

This paper presents a survey work on voice disability detection techniques available in the literature. It is shown in the literature that voice disability detection is a very challenging work, because the voice signal is very difficult 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 type of voice disability. Hence, it is very important to target a particular disability while designing the algorithm. In this survey paper, 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. In the survey, it is also found that a single letter, word or a full-sentence with pause can be used as voice samples. While using a full sentence some extra consideration should be given on transitional words and pause. 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, some extra precautions should be taken by the researchers as is mentioned in this survey paper. The selection of features from the samples is the next challenge for the researchers. From the survey, it is clear 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 a 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. Among these classification algorithms, SVM is considered the most suitable tool for voice disability detection. However, the SVM algorithm is not particularly suitable to categorize levels of voice disability. To achieve good accuracy in classification, a large data set is required to train the classifiers as well as to 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 features collection. Above all, they should focus on the design of a

level based voice pathology detection algorithm suitable for a distinct pathology.

APPENDIX

LIST OF ACRONYMS AND THEIR DEFINITIONS

AS	Asperger syndrome
ASR	Automatic speech recognition
APQ	Amplitude perturbation quotient
ANN	Artificial neural network
APR	The ratio of aperiodic/periodic components in cepstral energy
ACC	Classification accuracy
AR-HMM	Autoregressive higher-order HMM
ANOVA	Analysis of variance
BH	Bhattacharyya distance
CWT	Continuous wavelet transform
CHNR	Cepstral based HNR
CT	Computerized tomography
CNN	Convolutional neural network
CDBN	Convolutional deep belief network
CEP	Cepstral
CQ	Glottal closing quotient
CSL	Computerized speech laboratory
DWT	Discrete wavelet transform
DFT	Discrete Fourier transform
DCE	Delta cepstral
DNRNN	Dense Net Recurrent Neural Network
DBN	Deep belief network
DH	Degree of hoarseness
DPP	Dissimilarities in the surfaces of the pitch pulses
DUV	Degree of unvoiceness
EE	Emotional expression
EGG	Electroglottogram
EER	Equal error rate
FFT	Fast Fourier transform
FFNN	Feed forward neural network
GNR	Glottal to noise ratio
GMM	Gaussian mixture model
GMM-UBM	GMM-Universal background model
GMM-SVM	Gaussian Mixture Model Support Vector Machine
GRNN	Generalized regression neural network
GRBS	Grade Roughness Breathiness Strain
GNN	Graph neural network
HNR	Harmonic to noise ratio
HTK	Hidden Markov Model Tool Kit
HLAC	Higher-Order Local Autocorrelation
IQ	Intelligent quotient
IIR	All-pole infinite impulse response
JIT	Jitter
KL	Kullback Leibler
KL-MCS	Classical KL
LFCC	Linear frequency cepstral coefficient
LPC	Linear Predictive Coding
LP	Linear predictor
LEMG	Laryngeal electromyography
LDA	Linear discriminant analysis
MFCC	Mel-frequency Cepstral Coefficient
MRI	Magnetic resonance imaging
MEEI	Massachusetts Eye and Ear Infirmary
MLPNN	Multilayer perceptron neural network
MAP	Maximum a posterior probability algorithm
MDVP	Multi-dimensional Voice Program
NNR	Noise energy to total energy ratio
NNE	Normalized noise energy
NHR	Noise- to- Harmonic ratio
PLP	Perceptual linear prediction
PA	Pitch amplitude
PCA	Principal component analysis
PNN	Probabilistic neural network
PDF	Probability density function
PPQ	Pitch perturbation
PECM	Ratio of cepstral energies
RASTA-PLP	Relative spectral transform – PLP
RAP	Relative Average Perturbation
RNN	Recurrent Neural Network
RBM	Restricted Boltzmann machines
RBF	Radial basis function
RBFNN	Radial Basis Functional Neural Networks
SVD	Saarbruecken Voice Database
SVM	Support vector machine
SPI	Soft phonation index
SIR	Spectral Flatness of the Residue Signal
STFT	Short time Fourier transform
TEO	Teager energy operator
WCEP	Weighted cepstral
WDCEP	Weighted delta cepstral
ZCR	Zero crossing rate

REFERENCES

[1] N. Bhattacharyya, "The prevalence of voice problems among adults in the United States," *Laryngoscope*, vol. 124, no. 10, pp. 2359–2362, Oct. 2014, doi: 10.1002/lary.24740.

[2] M. A. Morris, S. K. Meier, J. M. Griffin, M. E. Branda, and S. M. Phelan, "Prevalence and etiologies of adult communication disabilities in the united states: Results from the 2012 national health interview survey," *Disability Health J.*, vol. 9, no. 1, pp. 140–144, Jan. 2016, doi: 10.1016/j.dhjo.2015.07.004.

[3] H. J. Hoffman, C. M. Li, K. E. Bainbridge, K. G. Losonczy, M. S. Chiu, and M. L. Rice, "Voice, speech, and language problems in the U.S. pediatric population: The 2012 national health interview survey (NHIS)," *Int. J. Epidemiology.*, vol. 44, p. i260, Oct. 2015, doi: 10.1093/ije/dyv096.489.

[4] American Speech-Language-Hearing. *Voice Disorders*. Accessed: Mar. 20, 2020. [Online]. Available: https://www.asha.org/practice-portal/clinical-topics/voice-disorders/

[5] T. E. Quateri, "Production and classification of speech sounds," in *Discrete-Time Speech Signal Processing: Principles and Practices*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001, pp. 72–76.

[6] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, K. H. Malki, T. A. Mesallam, and M. F. Ibrahim, "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," *IEEE Access*, vol. 6, pp. 6961–6974, 2018, doi: 10.1109/ACCESS.2017.2696056.

[7] L. Lee, L. G. Chamberlain, R. G. Loudon, and J. C. Stemple, "Speech segment durations produced by healthy and asthmatic subjects," *J. Speech Hearing Disorders*, vol. 53, no. 2, pp. 186–193, May 1988, doi: 10.1044/jshd.5302.186.
[8] A. Alzheimer, "On a peculiar disease of the cerebral cortex," *Allgemeine Zietschrift für Psychiatrie*, vol. 64, pp. 146–148, Jan. 1907.
[9] J. L. Cummings, D. F. Benson, M. A. Hill, and S. Read, "Aphasia in dementia of the alzheimer type," *Neurology*, vol. 35, no. 3, pp. 394–394, Mar. 1985, doi: 10.1212/wnl.35.3.394.
[10] K. Forbes and A. M. V. Shanks, "Distinct patterns of spontaneous speech deterioration: An early predictor of Alzheimer's disease," *Brain Congnition*, vol. 48, nos. 2–3, pp. 356–361, Mar./Apr. 2002, doi: 10.1006/brcg.2001.1377.
[11] V. Olga, B. Emery, and E. T. Oxman, *Dementia: Presentations, Differential Diagnosis, and Nosology* (The Johns Hopkins Series in Psychiatry and Neuroscience). Baltimore, MD, USA: The John Hopkins Univ. Press, 1994, pp. 108–122.
[12] A. Kertesz, J. Appell, and M. Fisman, "The dissolution of language in Alzheimer's disease," *Can. J. Neurol. Sci., J. Canadien des Sci. Neurologiques*, vol. 13, no. S4, pp. 415–418, Nov. 1986, doi: 10.1017/s031716710003701x.
[13] K. Faber-Langendoen, J. C. Morris, J. W. Knesevich, E. LaBarge, J. P. Miller, and L. Berg, "Aphasia in senile dementia of the alzheimer type," *Ann. Neurol.*, vol. 23, no. 4, pp. 365–370, Apr. 1988, doi: 10.1002/ana.410230409.
[14] L. Ferm, "Behavioural activities in demented geriatric patients," *Clinics*, vol. 16, no. 4, pp. 85–194, 1974, doi: 10.1159/000245521.
[15] H. S. Kirshner, "Progressive aphasia and other focal presentations of Alzheimer disease, Pick disease, and other degenerative disorders," *Dementia: Presentations, Differential Diagnosis, and Nosology*. Baltimore, MD, USA: Johns Hopkins Univ. Press, 1994, pp. 108–122.
[16] H. S. Kirshner, "Language disturbance: An initial symptom of cortical degenerations and dementia," *Arch. Neurol.*, vol. 41, no. 5, p. 491, May 1984, doi: 10.1001/archneur.1984.04050170037012.
[17] V. O. B. Emery, "Language impairment in dementia of the alzheimer type: A hierarchical decline?" *Int. J. Psychiatry Med.*, vol. 30, no. 2, pp. 145–164, Jun. 2000, doi: 10.2190/X09P-N7AU-UCHA-VW08.
[18] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *J. Neurol., Neurosurgery Psychiatry*, vol. 79, no. 4, pp. 368–376, Apr. 2008, doi: 10.1136/jnnp.2007.131045.
[19] R. A. Shirvan and E. Tahami, "Voice analysis for detecting Parkinson's disease using genetic algorithm and KNN classification method," in *Proc. 18th Iranian Conf. Biomed. Eng.*, Tehran, Iran, pp. 278–283, Dec. 2011, doi: 10.1109/ICBME.2011.6168572.
[20] K. M. Rosen, "Parametric quantitative acoustic analysis of conversation produced by speakers with dysarthria and healthy speakers," *J. Speech, Lang., Hearing Res.*, vol. 49, no. 2, pp. 395–411, Apr. 2006, doi: 10.1044/1092-4388(2006/031).
[21] B. Harel, M. Cannizzaro, and P. J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study," *Brain Cognition*, vol. 56, no. 1, pp. 24–29, Oct. 2004, doi: 10.1016/j.bandc.2004.05.002.
[22] P. A. LeWitt, "Parkinson's Disease: Etiologic Considerations," in *Parkinson's Disease and Movement Disorders*, C. H. Adler and J. E. Ahlskog, Eds. Totowa, NJ, USA: Humana Press, 2000, pp. 91–100, doi: 10.1007/978-1-59259-410-8_6.
[23] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Investigating the role of glottal features in classifying clinical depression," in *Proc. 25th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 3, Sep. 2003, pp. 2849–2852, doi: 10.1109/IEMBS.2003.1280512.
[24] M. Alpert, E. R. Pouget, and R. R. Silva, "Reflections of depression in acoustic measures of the patient's Speech," *J. Affect. Disorders*, vol. 66, no. 1, pp. 59–69, Sep. 2001, doi: 10.1016/s0165-0327(00)00335-9.
[25] Å. Nilsonne, J. Sundberg, S. Ternström, and A. Askenfelt, "Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression," *J. Acoust. Soc. Amer.*, vol. 83, no. 2, pp. 716–728, Feb. 1988, doi: 10.1121/1.396114.
[26] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 829–837, Jul. 2000, doi: 10.1109/10.846676.
[27] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," *J. Neurolinguistics*, vol. 20, no. 1, pp. 50–64, Jan. 2007, doi: 10.1016/j.jneuroling.2006.04.001.
[28] D. R. Weinberger, "Implications of normal brain development for the pathogenesis of schizophrenia," *Arch. Gen. Psychiatry*, vol. 44, no. 7, p. 660, Jul. 1987, doi: 10.1001/archpsyc.1987.01800190080012.
[29] B. Elvevåg, P. W. Foltz, M. Rosenstein, and L. E. DeLisi, "An automated method to analyze language use in patients with schizophrenia and their first-degree relatives," *J. Neurolinguistics*, vol. 23, no. 3, pp. 270–284, May 2010, doi: 10.1016/j.jneuroling.2009.05.002.
[30] J. Zhangi, "Clinical investigation of speech signal features among patients with schizophrenia," *Shanghai Arch. Psychiatry*, vol. 28, no. 2, pp. 95–102, Apr. 2016, doi: 10.11919/j.issn.1002-0829.216025.
[31] L. Kanner, "Irrelevant and metaphorical language in early infantile autism," *Amer. J. Psychiatry*, vol. 103, no. 2, pp. 242–246, Sep. 1946, doi: 10.1176/ajp.103.2.242.
[32] M. E. Hoque, "Exploring speech therapy games with children on the autism spectrum," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc.*, Brighton, U.K., Sep. 2009 pp. 1–4.
[33] S. E. Bryson, "Brief report: Epidemiology of autism," *J. Autism Develop. Disorders*, vol. 26, no. 2, pp. 165–167, Apr. 1996, doi: 10.1007/BF02172005.
[34] L. D. Shriberg, R. Paul, J. L. McSweeny, A. Klin, D. J. Cohen, and F. R. Volkmar, "Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome," *J. Speech, Lang., Hearing*, vol. 44, no. 5, pp. 1097–1115, Oct. 2001, doi: 10.1044/1092-4388(2001/087).
[35] P. Boersma, and D. Weenink. *PRAAT: Doing Phonetics by Computer, Version 6.1.10, Computer Program 2005*. Accessed: Apr. 4, 2020. [Online]. Available: http://www.fon.hum.uva.nl/praat/
[36] J. J. Diehl, D. Watson, L. Bennetto, J. Mcdonough, and C. Gunlogson, "An acoustic analysis of prosody in high-functioning autism," *Appl. Psycholinguistics*, vol. 30, no. 3, pp. 385–404, Jul. 2009, doi: 10.1017/S0142716409090201.
[37] A. Lacheret, M.-T. Le Normand, and S. Boushaba. (May 2008). *Prosodic Disturbances in Autistic Children Speaking French*. Campinas, Brazil. Accessed: Mar. 21, 2020. [Online]. Available: https://www.isca-speech.org/archive/sp2008/papers/sp08_195.pdf
[38] T. Oxman, "Diagnostic classification through content analysis of patient speech," *Amer. J. Psychiatry*, vol. 145, no. 4, pp. 464–468, Apr. 1988, doi: 10.1176/ajp.145.4.464.
[39] A. Maier, "Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer," *EURASIP J. Audio, Speech, Music Process.*, vol. 1, pp. 1–7, Dec. 2009, doi: 10.1155/.2010/926951.
[40] K. Graves, "Emotional expression and emotional recognition in breast cancer survivor," *J. Psychol. Health*, vol. 20, pp. 579–595, Jan. 2010, doi: 10.1080./0887044042000334742.
[41] *Voice Disorder*. John Hopkins Medicine. Accessed: Mar. 21,2020. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/voice-disorders
[42] V. Alan and R. W. Schafer, "Fourier transform and Fourier analysis of signals using the discrete Fourier transform," in *Discrete-Time Signal Processing*, 3rd ed. London, U.K.: Pearson, Aug. 2009, pp. 855–859.
[43] L. Rabiner and R. Schafer, "Algorithms for estimating speech parameter,"in *Theory and Application of Digital Speech Processing*, 1st ed. London, U.K.: Pearson, Mar. 2010, pp. 649–659.
[44] L. Tan and J. Jiang, "Subband and wavelet based coding," *Digital Signal Processing Fundamentals and Applications*, 2nd ed. New York, NY, USA: Academic, Oct. 2018, pp. 638–653.
[45] O. Buza, G. Toderean, A. Nica, and A. Caruntu, "Voice signal processing for speech synthesis," in *Proc. IEEE Int. Conf. Autom., Qual. Test., Robot.*, Cluj-Napoca, Romania, vol. 2, May 2006, pp. 360–364, doi: 10.1109/AQTR.2006.254660.
[46] D. O'Shaughnessy, "Linear predictive coding," *IEEE Potentials*, vol. 7, no. 1, pp. 29–32, Feb. 1988. [Online]. Available: https://ieeexplore.ieee.org/document/1890, doi: 10.1109/.45.1890.
[47] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990, doi: 10.1121/1.399423.
[48] D. J. Hermes. *Sound Perception: The Science of Sound Design*. Accessed: Mar. 21, 2020. [Online]. Available: http://home.ieis.tue.nl/dhermes/lectures/soundperception/04AuditoryFilter.html
[49] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994, doi: 10.1109/89.326616.

[50] J. P. Teixeira, D. Ferreira, and S. Carneiro. (2011). *Análise Acústica Vocal—Determinação do Jitter e Shimmer Para Diagnóstico de Patalogias da Fala*. 6° Congresso Luso-Moçambicano de Engenharia. Maputo, Moçambique. Accessed: Mar. 21, 2020. [Online]. Available: https://core.ac.uk/download/pdf/153407887.pdf

[51] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," Inst. Phonetic Sciences. Univ. Amsterdam, Amsterdam, The Netherlands, Tech. Rep., 1993, pp. 97–110, vol. 17. Accessed: Mar. 21, 2020. [Online]. Available: http://www.fon.hum.uva.nl/paul/papers/Proceedings_1993.pdf

[52] C. M. Vikram and K. Umarani, "Pathological voice analysis to detect Neurological Disorder using MFCC and SVM," *Int. J. Adv. Electr. Electron. Eng.*, vol. 2, no. 4, pp. 87–91, 2013.

[53] V. Srinivasan, V. Ramalingam, and P. Arulmozli, "Artificial neural network based pathological voice classification using MFCC features," *Int. J. Sci., Environ., Technol.*, vol. 3, no. 1, pp. 291–302, Feb. 2014. Accessed: Mar. 21, 2020. [Online]. Available: https://pdfs.semanticscholar.org/241b/313fd5758095d74abe8da7b8aa22e2348075.pdf

[54] H. Kasuya, S. Ogawa, and Y. Kikuchi, "An adaptive comb filtering method as applied to acoustic analyses of pathological voice," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Tokyo, Japan, vol. 11, Apr. 1986, pp. 669–672, doi: 10.1109/ICASSP.1986.1168996.

[55] G. D. Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *J. Speech, Lang., Hearing Res.*, vol. 36, no. 2, pp. 254–266, Apr. 1993, doi: 10.1044/jshr.3602.254.

[56] R. Cachu, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in *Proc. ASEE Regional Conf.*, Pittsburg, PA, USA, Jun. 2008, pp. 1–7. Accessed: Mar. 21, 2020. [Online]. Available: https://www.asee.org/documents/zones/zone1/2008/student/ASEE12008_0044_paper.pdf

[57] M. Slaney, "Auditory toolbox. version 2," Interval Research Corporation, Palo Alto, CA, USA, Tech. Rep. 1998-010, 1998. Accessed: Mar. 22, 2020. [Online]. Available: https://engineering.purdue.edu/~malcolm/interval/1998-010/AuditoryToolboxTechReport.pdf

[58] U. Shrawankar and V. M. Thakare, "Techniques for feature extraction in speech recognition system: A comparative study," *Int. J. Comput. Appl. Eng., Technol. Sci.*, pp. 412–418, May 2013. Accessed: Mar. 21, 2020. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1305/305.1145.pdf

[59] H. M. Teager, "Private communication," Dept. Biomed. Eng., Boston Univ. School Med., Boston, MA, USA, 1985.

[60] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Albuquerque, New Mexico, Apr. 1990, pp. 381–384, doi: 10.1109/ICASSP.1990.115702.

[61] J. F. Kaiser, "On Teager's Energy Algorithm and Its Generalization to Continuous Signals," in *Proc. 4th IEEE Digital Signal Process. Workshop*, New Paltz, NY, USA, Sep. 1990.

[62] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support vector clustering," *J. Mach. Learn. Res.*, vol. 2, pp. 125–137, Mar. 2002.

[63] *S. Gaussian-Mixture-Example*. Accessed: Mar. 22, 2020. [Online]. Available: https://commons.wikimedia.org/wiki/File:Gaussian-mixture-example.svg

[64] R. Douglas, *Universal Background Models*. Lexington, MA, USA: MIT Lincoln Laboratory, 2009, doi: 10.1007/978-0-387-73003-5_197.

[65] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, vol. 41, no. 1, pp. 164–171, Feb. 1970, doi: 10.1214/aoms/1177697196.

[66] L. E. Baum, "An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process," in *Proc. 3rd Symp. Inequalities*, Sep. 1969, pp. 1–8.

[67] R. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Mag.*, vol. 4, no. 2, pp. 4–22, Apr. 1987, doi: 10.1109/MASSP.1987.1165576.

[68] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015, doi: 10.1016/j.neunet.2014.09.003.

[69] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, Jan. 2009, doi: 10.1561/2200000006.

[70] Y. LeCun. *LeNet-5 Convolutional Neural Networks*. Accessed: Mar. 21, 2020. [Online]. Available: http://yann.lecun.com/exdb/lenet/

[71] Y. Zeinali and B. A. Story, "Competitive probabilistic neural network," *Integr. Comput.-Aided Eng.*, vol. 24, no. 2, pp. 105–118, Mar. 2017, doi: 10.3233/ICA-170540.

[72] G. Hinton, "Deep belief networks," *Scholarpedia*, vol. 4, no. 5, p. 5947, 2009, doi: 10.4249/scholarpedia.5947.

[73] D. F. Specht, "A general regression neural network," *IEEE Trans. Neural Netw.*, vol. 2, no. 6, pp. 568–576, Nov. 1991, doi: 10.1109/72.97934.

[74] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, nos. 2–3, pp. 131–163, Nov. 1997, doi: 10.1023/A:1007465528199.

[75] C. N. Silla and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Mining Knowl. Discovery*, vol. 22, nos. 1–2, pp. 31–72, Jan. 2011, doi: 10.1007/s10618-010-0175-9.

[76] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.* Berkeley, CA, USA: Univ. California Press, 1968, pp. 281–297.

[77] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.

[78] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. Hoboken, NJ, USA: Wiley, 2004.

[79] 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," *J. Voice*, vol. 33, no. 5, pp. 634–641, Sep. 2019, doi: 10.1016/j.jvoice.2018.02.003.

[80] C. M. Vikram and K. Umarani, "Phoneme independent pathological voice detection using wavelet based MFCCs, GMM-SVM hybrid classifier," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Mysore, India, Aug. 2013, pp. 929–934, doi: 10.1109/ICACCI.2013.6637301.

[81] F. Amala, M. Fezari, and H. Bourouba, "An Improved GMM-SVM system based on Distance Matrix for voice Pathology Detection," *Int. J. Appl. Math. Inf. Sci.*, vol. 10, no. 3, pp. 1061–1070, May 2016.

[82] A. Hossein, "A parametric approach for classification of distortions in pathological voice," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Process. (ICASSP)*, Calgary, Alberta, Apr. 2018, pp. 286–290, doi: 10.1109/ICASSP.2018.8461316.

[83] T. J. Jun and D. Kim. *Pathological Voice Disorders Classification From Acoustic Waveform*. Accessed: Mar. 21, 2020. [Online]. Available: http://mac.kaist.ac.kr/~juhan/gct634/2018/finals/pathological_voice_disorders_classification_from_acoustic_waveforms_poster.pdf

[84] D. Pravena, S. Dhivya, and D. Durga, "Pathological voice recognition for vocal cord disease," *Int. J. Comput. Appl.*, vol. 147, no. 13, pp. 31–37, Jun. 2012, doi: 10.5120/7250-0314.

[85] A. Mohammed, A. Mansour, M. Ghulam, Z. Mohammed, T. A. Mesallam, K. H. Malki, F. Mohamed, M. A. Mekhtiche, and B. Mohamed, "Automatic speech recognition of pathological voice," *Indian J. Sci. Technol.*, vol. 8, no. 32, pp. 1–6, Nov. 2015, doi: 10.17485/ijst/2015/v8i32/92130.

[86] M. A. Wahed, "Computer aided recognition of pathological voice," in *Proc. 31st Nat. Radio Sci. Conf. (NRSC)*, Apr. 2014, pp. 349–352, doi: 10.1109/NRSC.2014.6835096.

[87] S. C. Costa, B. G. A. Neto, J. M. Fechine, and S. Correia, "Parametric cepstral analysis for pathological voice assessment," in *Proc. ACM Symp. Appl. Comput. (SAC)*, Fortalex, Brazil, 2008, pp. 1410–1414, doi: 10.1145/1363686.1364011.

[88] M. Fezari, F. Amara, and I. El-Emary, "Acoustic analysis for detection of voice disorders using adaptive features and classifiers," in *Proc. Int. Conf. Circuits, Syst. Control*, Interlaken, Switzerland, Feb. 2014, pp. 112–117.

[89] J. Wang and C. Jo, "Vocal folds disorder detection using pattern recognition methods," in *Proc. 29th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Lym, France, Aug. 2007, pp. 3253–3256, doi: 10.1109/IEMBS.2007.4353023.

[90] L. Verde, G. De Pietro, and G. Sannino, "Voice disorder identification by using machine learning techniques," *IEEE Access*, vol. 6, pp. 16246–16255, 2018, doi: 10.1109/ACCESS.2018.2816338.

[91] J. R. Orozco-Arroyare, "Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 6, pp. 1820–1828, Nov. 2015, doi: 10.1109/JBHI.2015.2467375.

[92] D. Michaelis, M. Frolich, and H. W. Strube, "Selection and combination of acoustic features for the description of pathological voices," *J. Acoust. Soc. Amer.*, vol. 103, no. 3, pp. 1628–1639, Mar. 1998. [Online]. Available: https://asa.scitation.org/doi/10.1121/1.421305, doi: 10.1121/1.421305.

[93] A. Ali and S. Ganar, "Intelligent Pathological Voice Detection," *Int. J. Innov. Res. Technol.*, vol. 5, no. 5, pp. 92–95, Oct. 2018.

[94] M. Sarria-Paja, J. I. Daza-Santacoloma, G. Godino-Llorente, G. Castellanos-Domínguez, and N. Sáenz-Lechón, "Feature selection in pathological voice classification using dynamic of component analysis," in *Proc. 4th Int. Symp. Image/Video Commun. (ISVIC)*, Bilbao, Spain, Jul. 2008.

[95] V. Sellam and J. Jagadeesan, "Classification of normal and pathological voice using SVM and RBFNN," *J. Signal Inf. Process.*, vol. 5, no. 1, pp. 1–7, 2014, doi: 10.4236/jsip.2014.51001.
[96] M. Chopra, K. Khieu, and T. Liu. *Classification and Recognition of Stuttered Speech*. Stanford University. Accessed: Mar. 21, 2020. [Online]. Available: http://web.stanford.edu/class/cs224s/reports/Manu_Chopra.pdf
[97] R. J. Moran, R. B. Reilly, P. de Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 468–477, Mar. 2006, doi: 10.1109/TBME.2005.869776.
[98] Z. Kons, "On feature extraction for voice pathology detection from speech signals," in *Proc. 1st Annu. Afeka-AVIOS Speech Process. Conf.* Tel Aviv, Israel: Tel Aviv Academic College of Engineering, Jul. 2014.
[99] S. Bielamowicz, J. Kreiman, B. R. Gerratt, M. S. Dauer, and G. S. Berke, "Comparison of voice analysis systems for perturbation measurement," *J. Speech, Lang., Hearing Res.*, vol. 39, no. 1, pp. 126–134, Feb. 1996, doi: 10.1044/jshr.3901.126.
[100] L. Eskenazi, D. G. Childers, and D. M. Hicks, "Acoustic correlates voice quality," *J. Speech Hearing Res.*, vol. 33, pp. 298–306, Jun. 1990. [Online]. Available: https://pubs.asha.org/doi/10.1044/jshr.3302.298, doi: 10.1044/jshr.3302.298.
[101] A. Sasou, "Automatic identification of pathological voice quality based on the GRBAS categorization," in *Proc. APSIPA Annu. Summit Conf.*, Malaysia, India, Dec. 2017, pp. 1243–1247, doi: 10.1109/APSIPA.2017.8282229.
[102] B. Sabir, F. Rouda, Y. Khazri, B. Touri, and M. Moussetad, "Improved algorithm for pathological and normal voices identification," *Int. J. Electr. Comput. Eng.*, vol. 7, no. 1, pp. 238–243, Feb. 2017, doi: 10.11591/ijece.v7i1.pp238-243.
[103] S. Shinohara, Y. Omiya, M. Nakamura, N. Hagiwara, M. Higuchi, S. Mitsuyoshi, and S. Tokuno, "Multilingual evaluation of voice disability index using pitch rate," *Adv. Sci., Technol. Eng. Syst. J.*, vol. 2, no. 3, pp. 765–772, Jun. 2017, doi: 10.25046/aj020397.
[104] A. Dibazar, S. Narayan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Proc. 2nd Joint EMBS/BMES Conf.*, Houston, TX, USA, Oct. 2002, pp. 182–183, doi: 10.1109/IEMBS.2002.1134447.
[105] A. B. GUDI, H. K. Shreedhar, and H. C. Nagaraj, "Estimation of severity of speech disability through speech envelope," *Int. J. Signal Image Process.*, vol. 2, no. 2, pp. 26–33, Jun. 2011, doi: 10.5121/sipij.2011.2203.
[106] M. Sarria-Poja and G. Gastellones-Dominguaz, "Robust pathological voice detection based on component information from HMM," in *Proc. Int. Conf. Nonlinear Speech Process.*, 2011, pp. 254–261.
[107] A. H. Patil and V. P. Baljeker, "Classification of normal and pathological voices using TEO phase and Mel cepstral features," in *Proc. Int. Conf. Signal Process. Commun. (SPCOM)*. Bengalore, India: Indian Institute of Science, Jul. 2012, pp. 1–5, doi: 10.1109/SPCOM.2012.6289991.
[108] M. Alsulaiman, G. Mohammed, and Z. Ali, "Classifications of vocal fold disease using RASTA-PLP," in *Proc. Int. Conf. Bioinf. Comput. Biol.*, NV, USA, Mar. 2014.
[109] M. Redzuan, "An improved time domain pitch detection algorithm for pathological voice," *Amer. J. Appl. Sci.*, vol. 9, no. 1, pp. 93–102, Jan. 2012, doi: 10.3844/.ajassp.2012.93.102.
[110] B. Boynov, "A software system for pathological voice acoustic analysis," in *Proc. IMEKO Meas. Biol. Med.*, Dubrovrnik, Croatia, Sep. 1998, pp. 125–128.
[111] F. Perdigae, C. Nerves, and L. Sa, "Pathological voice detection using turbulent speech segments," in *Proc. Int. Conf. Bio-Inspired Syst. Signal Process.*, Vilamoura, Portugal, Feb. 2012, pp. 238–243, doi: 10.5220/0003775902380243.
[112] H. Wu *et al.*, "A deep learning method for pathological voice detection using convolutional deep belief network," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 446–450.
[113] P. Murphy, "Development of acoustic analysis techniques for use in diagnosis of vocal pathology," Ph.D. dissertation, School Phys. Sci., Dublin City Univ., Dublin, Republic of Ireland. Accessed: May 17, 2019. [Online]. Available: http://doras.dcu.ie/19122/1/Peter_Murphy_20130620152522.pdf
[114] G. Muhammad, M. Alsulaiman, A. Mahmood, and Z. Ali, "Automatic voice disorder classification using vowel formants," in *Proc. IEEE Int. Conf. Multimedia Expo*, Barcelona, Spain, Jul. 2011, pp. 1–6, doi: 10.1109/ICME.2011.6012187.
[115] V. Parsa and D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *J. Speech, Lang., Hearing Res.*, vol. 43, no. 2, pp. 469–485, Apr. 2000, doi: 10.1044/jslhr.4302.469.
[116] G. B. Kempster, "Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical procedure," *Amer. J. Speech-Language Pathol.*, vol. 18, no. 2, pp. 124–132, Mar. 2018, doi: 10.1044/1058-0360(2008/08-0017).
[117] *The Rainbow Passage: Detecting Vocal Cord Paralysis*. Accessed: Mar. 21, 2020. [Online]. Available: https://www. bergerhenryent.com/the-rainbow-passage-detecting-vocal-cord-paralysis/
[118] (1994). *Massachusette Eye & Ear Infirmry Voice & Speech LAB, Disordered Voice Database Model 4337 (Ver. 1.03)*. Kay Elemetrics Corporation. Lincoln Park, NJ, USA. Accessed: Mar. 22, 2020. [Online]. Available: https://www.masseyeandear.org/

RUMANA ISLAM received the M.Sc. degree in biomedical engineering from Wayne State University, Michigan, in 2004. She is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Windsor, Canada. She has 18 years of professional experience in industry and academy. She served as a Senior Engineer at SIMENS. She worked as a Lecturer at the Department of Electrical and Electronic Engineering, American International University, Bangladesh, from 2007 to 2011. She is currently working as an Adjunct Faculty Member with the Department of Electrical Engineering, University of Science and Technology of Fujairah, UAE. She has published 15 research articles in highly reputed journals and conference proceedings. Her research interests are in biomedical engineering, biomedical signal processing, digital signal processing, and biomedical image processing.

MOHAMMED TARIQUE received the Ph.D. degree in electrical engineering from the University of Windsor, Canada, in 2007. He had worked as an Assistant Professor at the Department of Electrical Engineering, American International University, Bangladesh, from 2007 to 2011. He joined Ajman University, in 2011. He is currently working as an Associate Professor of electrical engineering with the University of Science and Technology of Fujairah (USTF). His research interests are in wireless communication, mobile ad hoc networks, digital communication, digital signal processing, and biomedical signal processing. He has published 35 research articles in highly reputed journals on the above mentioned fields. He also published 15 research articles in the proceedings of highly reputed international conferences.

ESAM ABDEL-RAHEEM (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from Ain Shams University, Cairo, Egypt, in 1984 and 1989, respectively, and the Ph.D. degree from the University of Victoria, Canada, in 1995, all in electrical engineering. In 1997, he joined Ain Shams University as an Assistant Professor. From 1999 to 2001, he was a Senior Design Engineer with AMD, Sunnyvale, CA, USA. He was an Adjunct Associate Professor with the University of Victoria, BC, Canada, from 2003 to 2009. Since 2003, he has been with the ECE Department, University of Windsor, ON, Canada, where he is currently a Full Professor. He has authored or coauthored about 110 articles in refereed journal and conference publications, including one published U.S./EU/world patent. His research interests include digital signal/image/video processing, signal processing for communications, VLSI signal processing systems, and spectrum sensing for cognitive radio networks. He was an Editorial Broad Member of *IET Signal Processing*, from 2006 to 2018. He is a member of the Association of Professional Engineers Ontario. He has served as the Technical Program Co-Chair for IEEE ISSPIT 2004 and 2005 and EIT 2009. He was the General Co-Chair for IEEE ISSPIT 2007, IEEE CCECE'2017, and IEEE ISSPIT 2018. He served as an Associate Editor for *Canadian Journal of Electrical and Computer Engineering*, from 2006 to 2010, and *Circuits, Systems and Signal Processing*, from 2007 to 2015.