Sound & Vibration

Tech Science Press

DOI: 10.32604/sv.2021.011734

ARTICLE

Combined Signal Processing Based Techniques and Feed Forward Neural Networks for Pathological Voice Detection and Classification

T. Jayasree[1,*] and S. Emerald Shia[2]

[1]Department of Electronics and Communication Engineering, Government College of Engineering, Tamil Nadu, 629007, India

[2]Department of Electronics and Communication Engineering, Cape Institute of Technology, Tamil Nadu, 629001, India

[*]Corresponding Author: T. Jayasree. Email: jayasree@gcetly.ac.in

Received: 26 May 2020 Accepted: 13 August 2020

ABSTRACT

This paper presents the pathological voice detection and classification techniques using signal processing based methodologies and Feed Forward Neural Networks (FFNN). The important pathological voices such as Autism Spectrum Disorder (ASD) and Down Syndrome (DS) are considered for analysis. These pathological voices are known to manifest in different ways in the speech of children and adults. Therefore, it is possible to discriminate ASD and DS children from normal ones using the acoustic features extracted from the speech of these subjects. The important attributes hidden in the pathological voices are extracted by applying different signal processing techniques. In this work, three group of feature vectors such as perturbation measures, noise parameters and spectral-cepstral modeling are derived from the signals. The detection and classification is done by means of Feed Forward Neural Network (FFNN) classifier trained with Scaled Conjugate Gradient (SCG) algorithm. The performance of the network is evaluated by finding various performance metrics and the the experimental results clearly demonstrate that the proposed method gives better performance compared with other methods discussed in the literature.

KEYWORDS

Autism spectrum disorder; down syndrome; feed forward neural network; perturbation measures; noise parameters; cepstral features

1 Introduction

Voice disability is a barrier to communication involving speech, hearing, language and fluency. In the World's total population, about 1.2% is facing some type of voice disability. Different types of surgical procedures and medical tests are used for the diagnosis of voice disability diagnosis [1]. For the past decades, many researchers have been working to find alternative methods to the conventional surgical procedures and medical tests. Voice sample based diagnosis is one of them. In this method, first the voice samples are extracted from the persons using different signal processing techniques. Then, the pathological voices are classified from the normal voices using distinct classification approaches.

In this paper, two important voice disabilities are considered, namely: Autism or Autism Spectrum Disorder (ASD) and Down Syndrome (DS). Autism is a complex neurological developmental disorder that affects a person's ability to communicate and interact with others. The signs of autism typically appear during early childhood. As there are many different indications of autism and the symptoms can

be from mild to severe, it is often referred to as Autism Spectrum Disorder (ASD). There are approximately one in 59 children in the United States has been identified with ASD [2] and 23 of every 10,000 children in India have autism [3]. Autism manifests itself in different ways in children and adults. The speech of children with ASD appears abnormal and is described as machine-like "monotonic" or "sing-song" [4]. Down syndrome (DS) is the chromosomal disorder caused by the presence of a third copy of chromosome 21. The phenotypic characteristics of DS include mental retardation, general hypotonia (decreased muscle tone), maxillary hypoplasia (underdevelopment of maxillary bone) with a relative macroglossia (unusually large tongue), short neck, and obesity. All of this can contribute to particular acoustic alteration [5]. Down Syndrome occurs in about one out of every 700 babies born in the US [6] and in India, the reported incidence of Down syndrome is one in 1250 [7].

Normally, specially trained physicians and psychologists diagnose ASD using ASD-screening tools like Autism Diagnostic Observation Schedule Revised (ADOS-R) and Autism Diagnostic Interview-Revised (ADI-R) [4]. But these methods require standardized structures for capturing the behaviors and found to be very difficult. Similarly, Down Syndrome Screening tests are used for diagnosing Down Syndrome. Nevertheless, they cannot give appropriate results [5]. There are many challenges and issues related to voice signal based pathology detection and classification techniques. Some important concerns are the selection of suitable voice features and selection of appropriate classifiers.

The perceptual voice quality of the majority of the speakers with Autism and Down Syndrome exhibits breathiness, roughness, hoarseness, and is low pitched. Although several studies have analyzed the variability of acoustic features in the voice produced by children with ADS, DS and normal children, there has been no attempt to classify both. Signal processing techniques are introduced and they are found to be more effective for the analysis of voice pathology signals.

The identification of pitch or fundamental frequency is crucial for the analysis of voice signals. The fundamental frequency (f_0) or pitch resembles perceptually to the total number of times per second the vocal folds come together for the entire duration of phonation [6]. Besides, voice perturbation measures such as jitter and shimmer based on the fundamental frequency are the other two significant measures used for the extraction of features from the voice disorder signals. Jitter deals with varying loudness in the voice whereas, shimmer deals with a recurrent back and forth variation in amplitude in the voice [7].

Albertini et al. [8] suggested jitter and shimmer for detecting the Down Syndrome and attained higher mean f_0 and lower spectral energy for the adults who were affected with Down Syndrome. Besides, they also showed that there is no marked difference between the voice characteristics of children with and without Down Syndrome. Lee et al. [9] utilized the measures of phonation in continuous speech and showed that the speakers with DS exhibited higher mean f_0, reduced pitch range, reduced jitter and attained no significant deviation in shimmer compared to the controls. Moura et al. [10] focused the performance measures such as jitter and shimmer in vowels acquired from the children with Down Syndrome and ascertained that the children with Down Syndrome produced lower f_0 with high dispersion than normal children for all the five vowels. Jeffery et al. [11] investigated the sustained vowels from four young adults with Down Syndrome for f_0, jitter, shimmer and showed that intermittent subharmonics were evident in spectrograms, some of which coincides with perceived diplophonia. The main limitation of voice perturbation measures is that a high degree of jitter consequences in a voice with roughness which is commonly perceived in the recordings of pathological voices. Besides, it is very difficult to measure fundamental frequency, in the case of a pathological voice [12]. This leads to the generation another set of features based on the noise energy present in the signal.

The Harmonic-to-Noise Ratio (HNR), Glottal-to-Noise Excitation Ratio (GNE) and Signal to Noise Ratio (SNR) are the noise energy based features used for voice pathology analysis. Sampaio et al. [13] projected the noise parameters for the diagnosis of diseases in dysphonic voices. The HNR is calculated

in dB as the average difference between the harmonic peaks and the aperiodic components of the signal [14]. Hamid et al. [15] employed EMPEG-7 feature set consisting of noise energy based measures for the early detection of autism. The main problem faced in these approaches is that, the noise parameters depend on the voice recording environment. Hence the values of parameters also vary accordingly. In order to overcome the limitations of the aforementioned feature extraction methods, the important attributes hidden in the voice pathology signals are found out by analyzing the spectral and cepstral characteristics of the signals.

The spectral features are obtained by finding the Perceptual Linear Prediction (PLP) models of the signals. These models are based on the concept of the psychophysics of hearing are used for the analysis of the voice pathology signals. The important application of the PLP is to remove irrelevant information confined in the speech [16]. Besides, PLP has spectral characteristics that are transformed to match the human auditory system. Another prevalent feature used in voice disability detection is RASTA-PLP [17]. A special bandpass filter known as called RASTA filter is employed in computing the RASTA-PLP.

More information can be viewed from the Mel-frequency cepstrum of the signal. It is the depiction of a sound signal expressed in terms of the linear cosine transform of a log power spectrum on a nonlinear mel-scale of frequency. Consequently, the coefficients that jointly encompass the Mel-frequency Cepstrum are called MFCC features. One of the important characteristics of MFCC is that, they try to analyze the vocal tract independently of the vocal folds that can be injured due to voice pathologies [18]. Yoram et al. [19] used long term frequency spectrum based features for detecting the autism disorder using autistic and normal speech samples. Deng et al. [14] adopted Extended Geneva Minimalistic Acoustic Parameter Set (EGEMAPS) and Computational Paralinguistics Challenge Feature set (COMPARE) consisting of spectral and cepstral based features.

After the feature extraction stage, the detection/classification task will begin. Numerous binary and multi-classification methods have been used for classification and detection. Support Vector Machines (SVM) are used for nonlinear regression and pattern classification. In the SVM approach, the low-dimensional training data is projected in a higher dimensional feature space. The patterns for training the SVM were attained from the recordings of children voices, both for normal and pathological [16]. Dankovicova et al. [20] used K-Nearest Neighbor (KNN) classifier for recognizing the pathological speech with the significant improvement in classification accuracy. Navie Bayes classifier based on the probability model is also used for classification purposes [21].

Moreover, it is also found that, most of the existing voice based-autism detection techniques use either short speech utterances or utterances segmented from spontaneous speech to detect autism. This work proposes automatic voice pathology detection using signal processing techniques with the sustained vowel /ah/. It employs a Feed Forward Neural Network (FFNN) to perform the classification. In the case of voice-based DS analysis, the existing systems have studied the variability of the acoustic features in the speech of DS and normal subjects employing the measures of central tendency such as mean, SD, range and dispersion but there has been no attempt to classify both.

This paper presents a new approach that aims to discriminate DS subjects from normal ones using a Feed-Forward Neural Network (FFNN), which to the best of our knowledge has not been shown before. This study also accomplishes the classification between ASD and DS using the sustained vowel /ah/. Moreover, this work also presents a comparative evaluation of the performance of our proposed systems with three different acoustic feature sets including perturbation measures, noise parameters and, cepstral features. The rest of the paper is organized as follows. Section 2 covers the methodology used in this work, Section 3 explains the experiments and Section 4 presents the results and discussion. The final section presents the conclusion.

2 Proposed Methodology

This work proposes, an automated system based on signal processing and artificial intelligence to detect and classify ASD and DS voices. The system model consists of three stages such as preprocessing and feature extraction followed by classification stage. The basic blocks of the method employed in this study are shown in Fig. 1.

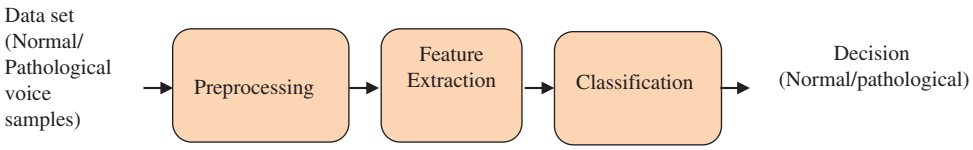

Figure 1: Basic blocks of the method employed

2.1 Preprocessing

The speech signal should be preprocessed to extract the relevant features. The steps involved in preprocessing are pre emphasis, framing, and windowing.

Pre emphasis: In the frequency spectrum of voiced speech, the high-frequency components have greater magnitude compared to the low frequencies. Pre emphasis is done to spectrally flatten the signal and to avoid numerical problems during the Fourier Transform operation [20]. Pre emphasis is achieved by passing the digitized speech signal $x(n)$ at time n through a low order digital system whose output $\tilde{x}(n)$ is related to input $x(n)$ by the difference equation

$$\tilde{x}(n) = x(n) - kx(n-1) \tag{1}$$

where $k = 0.97$ is the pre emphasis coefficient.

Here K is the pre emphasis coefficient where $K \in (0.9, 1.0)$. The typical value of k is 0.97. This entails that the difference between the current sample *x(n)* and the previous sample *x (n - 1)* is very less. This smaller difference ensures high pass filtering, because the difference between the consecutive samples is high only for high frequency components.

Framing: Normally a speech signal is not stationary. But when examined over a sufficiently short period of time its characteristics are stationary. So the pre emphasised speech signals $\tilde{x}(n)$ is divided into N frames of L samples with an overlap of D samples between adjacent frames. Then the i^{th} frame is given by the expression

$$x_i(n) = x(n + iD); n = 0, 1, 2 \ldots L - 1,\ i = 0, 1, 2 \ldots N - 1 \tag{2}$$

By introducing overlap, the transition between frames is reduced. A long window is advisable in order to attain good frequency in signal resolution, but the significance of some short transmission makes a short window more appropriate and effective. A common negotiation in the quality of the signal, that is always accessible to patch up if the signal frame length is about 20 or 30 ms, and with a frame spacing of 5 to 15 ms [21].

Windowing: Each frame is then multiplied by a hamming window to enhance the harmonics and to eliminate the discontinuities at the edges for the subsequent power spectrum computation. If we denote the window as $w(n), 0 \leq n \leq N - 1$. Then the result of windowing is

$$\tilde{x}_i(n) = x_i(n)w(n), \qquad 0 \leq n \leq N - 1 \tag{3}$$

The common hamming window used in this work has the form

$$w(n) = \begin{cases} 0.54 + 0.46\cos\left(\dfrac{2\pi n}{N-1}\right); 0 \le n \le N-1 \\ 0; otherwise \end{cases} \quad (4)$$

In hamming window, the width of the main lobe is greater than that of the other windows and it also has lower sidelobe amplitudes. Adding to its advantage, the approximate difference between pass band and stop band gains is about 43 dB. Hamming window is usually preferred because it generates less oscillation in the side lobes.

2.2 Feature Extraction

Fig. 2 explains the detailed description of the methodology applied in this work for comparative evaluation of the three proposed systems.

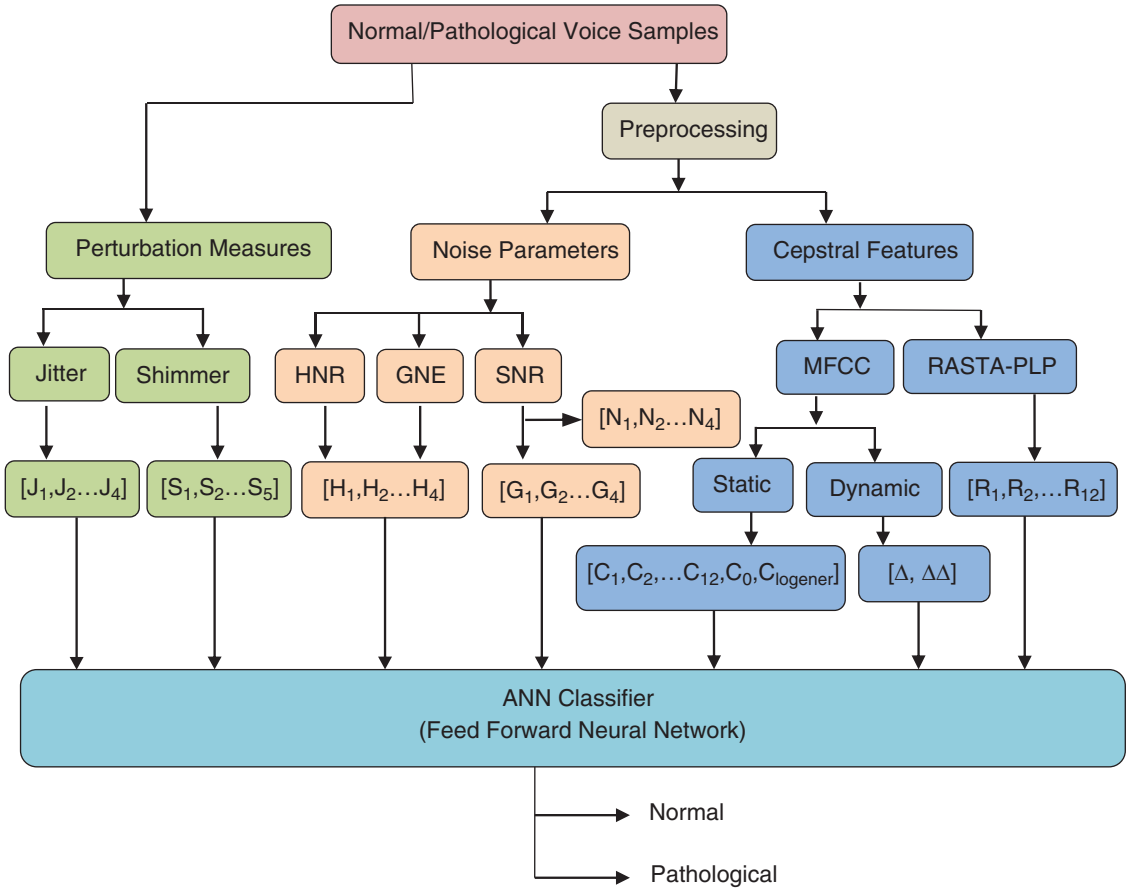

Figure 2: Block diagram for speech signal based detection and classification of autism and down syndrome

2.2.1 Perturbation Measures

The fundamental frequency is the rate of vibration of vocal folds. The vibratory cycles of an abnormal voice are more erratic compared to the voice produced by a normal person. The Figs. 3a–3c show the voiced signals acquired from normal, ADS and DS children for the sustained vowel /ah/.

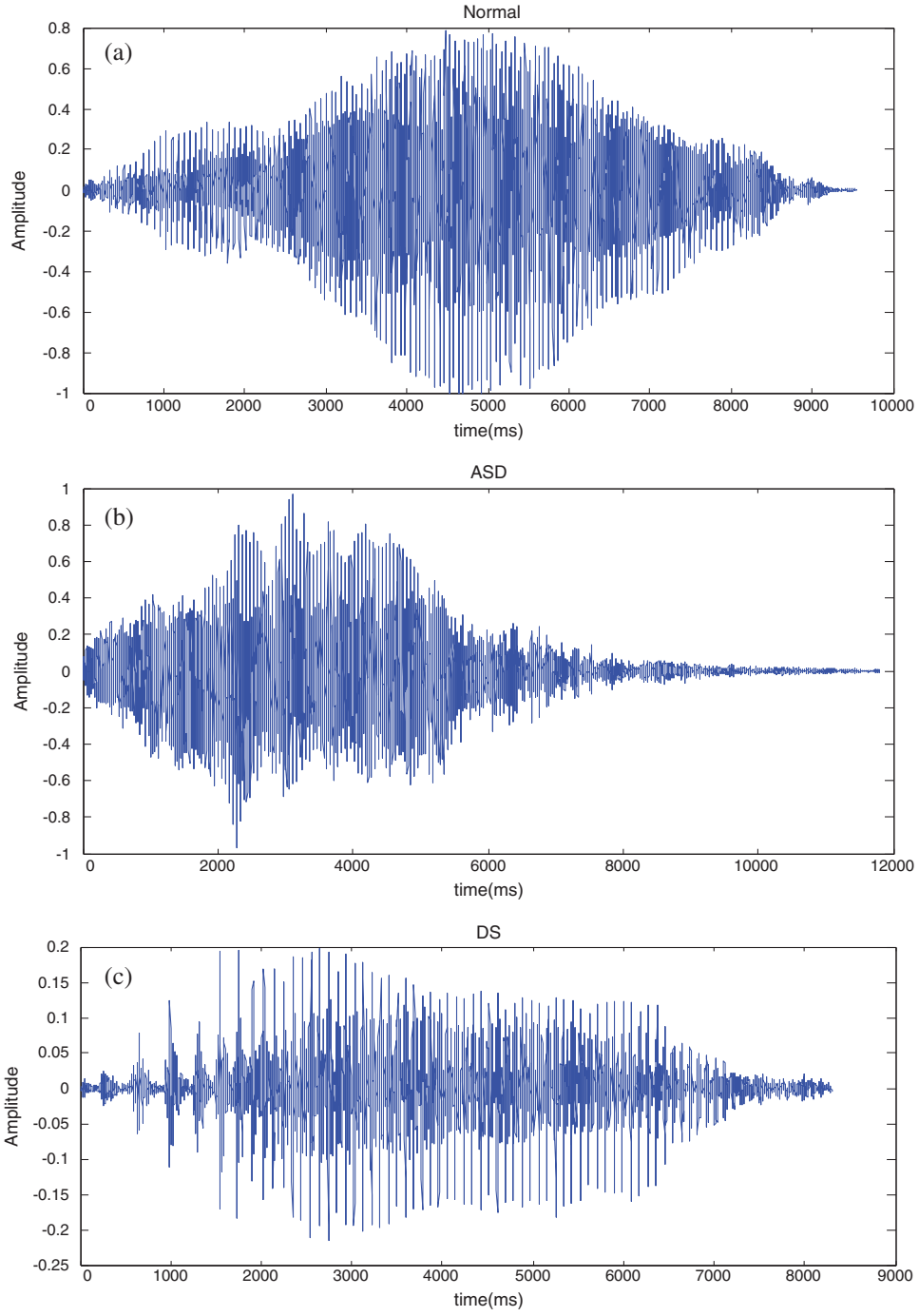

Figure 3: Speech signals acquired from Normal, ASD and DS children (a) Speech signal for normal (b) Speech signal for Autism Spectrum Disorder (ASD) (c) Speech signal for Down Syndrome (DS)

The perturbation measures jitter and shimmer quantify the cycle-to-cycle variability in f_0 and amplitude respectively of a speech signal. The four related jitter parameters namely absolute jitter ($Jitter_{(abs)}$), local jitter ($Jitter_{(loc)}$), three-point pitch perturbation quotient ($Jitter_{(PPQ3)}$) and five-point pitch perturbation quotient

(Jitter$_{(PPQ5)}$) are computed from f_0 contours obtained by SHRP algorithm [21]. If N is the number of f_0 computations, then the jitter variants are given as [22]

$$\text{Jitter}_{(abs)} = \frac{1}{N}\sum_{i=1}^{N-1}\left|F_{0(i)} - F_{0(i+1)}\right| \tag{5}$$

$$\text{Jitter}_{(loc)} = \frac{\frac{1}{N}\sum_{i=1}^{N-1}\left|F_{0(i)} - F_{0(i+1)}\right|}{\frac{1}{N}\sum_{i=1}^{N} F_{0(i)}}.100 \tag{6}$$

$$\text{Jitter}_{PPQ3} = \frac{\frac{1}{N-2}\sum_{i=1}^{N-1}\left|F_{0(i)} - \left(\frac{1}{3}\sum_{n=i-1}^{i+1} F_{0(n)}\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} F_{0(i)}} \tag{7}$$

$$\text{Jitter}_{PPQ5} = \frac{\frac{1}{N-2}\sum_{i=3}^{N-2}\left|F_{0(i)} - \left(\frac{1}{5}\sum_{n=i-2}^{i+2} F_{0(n)}\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} F_{0(i)}} \tag{8}$$

Shimmer is the analog of jitter, for the amplitude of speech signal. It can be obtained from the amplitude contours a_0 instead of f_0 contours. For computation of a_0 contours first the glottal cycles are obtained using DYPSA [23] algorithm. Then A_0 contour is the maximum amplitude value within each glottal cycle. The various shimmer parameters are

$$\text{Shimmer}_{(abs)} = \frac{1}{N}\sum_{i=1}^{N-1}\left|A_{0(i)} - A_{0(i+1)}\right| \tag{9}$$

$$\text{Shimmer}_{(loc)} = \frac{\frac{1}{N}\sum_{i=1}^{N-1}\left|A_{0(i)} - A_{0(i+1)}\right|}{\frac{1}{N}\sum_{i=1}^{N} A_{0(i)}}.100 \tag{10}$$

$$\text{Shimmer}_{PPQ3} = \frac{\frac{1}{N-2}\sum_{i=1}^{N-1}\left|A_{0(i)} - \left(\frac{1}{3}\sum_{n=i-1}^{i+1} A_{0(n)}\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} A_{0(i)}} \tag{11}$$

$$\text{Shimmer}_{PPQ5} = \frac{\frac{1}{N-2}\sum_{i=3}^{N-2}\left|A_{0(i)} - \left(\frac{1}{5}\sum_{n=i-2}^{i+2} A_{0(n)}\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} A_{0(i)}} \tag{12}$$

$$\text{Shimmer}_{(dB)} = \frac{1}{N}\sum_{i=1}^{N-1} 20\log\left|\left(\frac{A_{0(i)}}{A_{0(i+1)}}\right)\right| \tag{13}$$

Jitter and shimmer have been successfully used in early detection of articular pathology, Parkinson's disease and vocal fold disorders [24].

2.2.2 Noise Parameters

The measures Harmonic to Noise Ratio (HNR), Signal to Noise Ratio (SNR) and Glottal Noise Excitation Ratio (GNE) evaluate the noise content in the speech signal. For a speech signal *x(t)* with harmonic component *h(t)* and noise component n(t), the HNR is computed in terms of normalized autocorrelation ($R'_{xx}(\tau)$) at maximum lag τ_{max} according to the formula [25]:

$$HNR_{dB} = 10 \log\left(\frac{R'_{xx}(\tau_{max})}{1 - R'_{xx}(\tau_{max})}\right) \tag{14a}$$

where, $R'_{xx}(\tau_{max}) = \frac{R_{hh}(0)}{R_{xx}(0)}$ is the relative power of harmonic component and its complement $1 - R'_{xx}(\tau_{max}) = \frac{R_{nn}(0)}{R_{xx}(0)}$ is the relative power of the noise component in which $R_{hh}(0)$, $R_{nn}(0)$ represent autocorrelation of harmonic component and noise component where as $R_{xx}(0)$ represents the speech signal at zero lag.

The Signal to Noise Ratio is calculated using the formula

$$SNR = 20\log(S/N) \tag{14b}$$

where S is the signal power and N is the noise power

The GNE computation relies on correlation between Hilbert envelopes of different frequency channels, distributed throughout the speech spectrum. For normal speech during each glottis closure, all the frequency channels are simultaneously excited so that the Hilbert envelopes in all the channels have the same shape, leading to high correlation between them. In the case of noisy signals (Breathy speech), narrowband noise excites each frequency channel in a different manner, reducing the correlation between envelopes [26]. The steps involved in computation of GNE parameters are

1) Down sample the speech signal to 10 KHz.
2) Inverse filter the speech signal to detect the glottal cycles.
3) Compute the Hilbert envelope of different frequency bands using a specified bandwidth for each glottal cycle.
4) Calculate the cross correlation function for each pair of envelopes. Choose the pair of envelopes such that the difference between their center frequencies is equal to or greater than Half Bandwidth.
5) Pick the maximum of each correlation function.
6) Choose the maximum of Step 5, which is the GNE value for the detected glottal cycle.
7) Compute the mean, SD, Skewness, Kurtosis of GNE values for different glottal cycles.

2.2.3 Cepstral Features

(i) Mel Frequency Cepstral Coefficients (MFCCs): MFCCs have been successfully used in the detection of neurological diseases such as parkinson's disease [22], laryngeal pathologies, and hyper nasality associated with cleft lip and palate [27]. As a first step to compute MFCC the windowed frames $\tilde{x}_i(n)$ are applied with FFT to obtain the short term power spectrum. The FFT $X(k)$ and hence the power spectrum $P(k)$ of the speech signal are given as

$$X(k) = \sum_{n=0}^{N-1} \tilde{x}_i(n) e^{\frac{-j2\pi kn}{N}} \quad , \ 0 \ \leq k < N \tag{15}$$

$$P(k) = |X(k)|^2 = Re^2[X(k)] + Im^2[X(k)] \tag{16}$$

The power spectrum gives information about the amount of energy contained in each frequency band. MFCC extraction is a method motivated by the behavior of human auditory system. The human auditory system perceives sound in a non-linear frequency binning. The nonlinear signal processing characteristic and the spectral filtering behavior of inner ear to sound stimulus can be simulated by the Mel-bank filtering procedure [28]. If the simulated Mel-filter bank is $H_m[k]$. The Mel-frequency spectrum S[m] is obtained by filtering the speech signal spectrum $|X(k)|^2$ with the Mel-filter bank.

$$S[m] = \left[\sum_{k=0}^{N-1} |X(k)|^2 H_m[k] \right] \quad ; 0 < m \leq M \tag{17}$$

The Mel-frequency cepstrum $c[n]$ is then computed as the discrete cosine transform of log of Mel-frequency spectrum.

$$\mathrm{c[n]} = \sum_{\mathrm{m=o}}^{\mathrm{M-1}} \ln(S[\mathrm{m}]) \cos\left(\frac{\pi \mathrm{n}\left(\mathrm{m} - \frac{1}{2}\right)}{\mathrm{M}} \right) \quad ; 0 \leq n < M \tag{18}$$

The value of M varies between 20–40 for different applications. The value of M chosen for our implementation is 40.

Dynamic MFCCs: Dynamic MFCC coefficients called as delta (Δ) and delta-delta (ΔΔ) coefficients are the first and second derivatives of the obtained static MFCC features. The first temporal derivative Δ (referred to as differential coefficients) can be computed from static MFCCs [27] as

$$\Delta_t^i = \frac{\sum_{\theta=1}^{\Theta} \theta\left(c_{t+\theta}^i - c_{t-\theta}^i\right)}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{19}$$

where Δ_t^i denotes the i^{th} delta coefficient of the t^{th} frame and c is the static MFCC parameter.

The second temporal derivative ΔΔ can computed from Δ features according to the formula

$$\Delta\Delta_t^i = \frac{\sum_{\theta=1}^{\Theta} \theta\left(\Delta_{t+\theta}^i - \Delta_{t-\theta}^i\right)}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{20}$$

The typical value of Θ is 2. ΔΔ parameters are also called as acceleration coefficients. Static MFCC extracts only the static information of the acoustic signal spectrum whereas the dynamic parameters represent the inter-frame variations.

(ii) Rel Ative SpeTtrAl-Perceptual Linear Predictive (RASTA-PLP): RASTA-PLP was developed by Hermansky et al. [29]. It is an improvement over PLP. PLP's goal is to describe the psychophysics of human hearing more accurately in the feature extraction process. RASTA filtering is incorporated with PLP to eliminate the effect of non-speech components such as communication channel influence in the speech signal. The steps involved in the computation of RASTA-PLP features are [29,30]:

1) For each frame compute the power spectrum.

2) Compute the critical band power spectrum using bark spaced filter bank.

3) Dynamically compress the spectral amplitude by applying natural logarithm to the critical band power spectrum.

4) RASTA filter the compressed critical band spectrum to eliminate the effect of constant and slowly varying components introduced in speech by the communication channel.

5) Apply inverse log to the output of RASTA filter.

6) Multiply the resulting critical band spectrum by the equal loudness curve and raise it to the power of 0.33 to simulate the power law of hearing.

7) Compute the all-pole model of the resulting spectrum to extract the RASTA-PLP features.

3 Classification Using Feed Forward Neural Network (FFNN)

The FNN consists of an input layer, a single hidden layer and an output layer as shown in Fig. 4. The size of the input layer is equal to the size of input feature vector and the size of the output layer is equal to the number of target classes to be distinguished. The hidden layer may have varying number of neurons. The number of hidden neurons that produce best results are fixed experimentally. The three layer FFNN shown in Fig. 4 receives inputs $x_1, x_2, x_3 \ldots x_n$ processes them and forwards them to the hidden layer and then to the output layer to give the outputs $y_1, y_2, y_3 \ldots y_q$. The outputs of hidden layer are $z_1, z_2, z_3 \ldots z_p$. The weight w_{ij} connects the input node i to the hidden node j and the weight w_{jk} connects the hidden node j to the output node k.

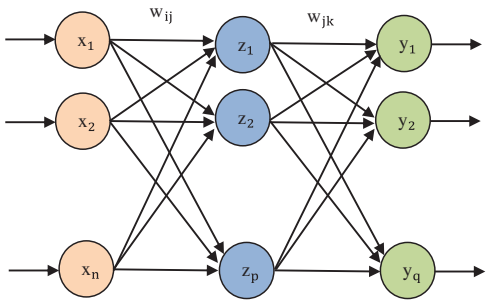

Figure 4: Feed forward neural network architecture

For each pattern h, $z_j(h)$ and $y_k(h)$ are the output of hidden layer and output layer are respectively and are given as [30]

$$z_j(h) = f\left(\sum_{i=1}^{n} x_i(h) w_{ij}\right) \qquad i = 1, 2, 3 \ldots n \; ; \; j = 1, 2, 3 \ldots p \tag{21}$$

$$y_k(h) = g\left(\sum_{j=1}^{p} z_j(h) w_{jk}\right) \qquad j = 1, 2, 3 \ldots p; \;\; k = 1, 2, 3 \ldots q \tag{22}$$

where, f and g are the activation functions. The most commonly used activation functions are sigmoidal and hyper-tangent. For binary classification q = 2.

Let the target output be t_1, t_2, t_3….t_q. The training objective is to determine the weight values that minimize the difference between the desired output and actual network output for all the patterns. So, the error criterion can be written as

$$E(h) = \frac{1}{H}\sum_{h=1}^{H}\sum_{k=1}^{p}(t_k(h) - y_k(h))^2 \tag{23}$$

where h represents pattern number and k represents output node number.

The proposed FFNN uses a supervised learning algorithm called Scaled Conjugate Gradient (SCG) algorithm [31–33] to update weights. SCG combines Levenberg-Marqardt algorithm with conjugate gradient approach to solve the weight vector optimization problem. The SCG update weights recursively as [32]

$$v_{ij}(iter+1) = v_{ij}(iter) + \alpha(iter)\rho(iter) \tag{24}$$

$$w_{jk}(iter+1) = w_{jk}(iter) + \alpha(iter)\rho(iter) \tag{25}$$

Here $\alpha(iter)$ and $\rho(iter)$ are the search direction and step size respectively in the specified iteration and are determined as explained in Møller [32]. The SCG is more efficient than other conventional Back Propagation algorithm. Since, the SCG does not contain any user dependent parameters, whose values are crucial for its success also a step size scaling mechanism is used to eliminate time consuming line search per iteration [34].

The important objective of the FFNN is to develop a model that performs well both on the training and the testing dataset on which the model would be used to make estimates. The ability of the network to perform better on the previously unseen inputs is called generalization. Hence, it is necessary to develop a model that can learn from the known examples and generalize from those examples to new examples in the future. The k-fold validation approach is used for estimating the ability of the model for generalizing the new data. However, learning and generalization is very difficult. If the learning is very less, the performance of the network is poor and if the learning is high, the model will perform well on the training dataset and poorly on new data, hence the model will overfit the problem. Thus, it is necessary to develop a model that suitably learns the training dataset and generalizes well to the new dataset. For better generalization, it is necessary to approximate the target function.

The phenomenon of underfitting can be addressed by increasing the capacity of the model. The capacity indicates the capability of the model to fit a range of functions. The capacity of the model can be improved by varying the structure of the model, such as adding more layers and/or more nodes to the layers. An overfit model is identified by observing the performance of the model during training by evaluating it on both training and validation dataset.

Another important issue in the design of neural network is the fascination of hidden neurons with minimum error and highest accuracy. The training set and generalization error are expected to be high before learning starts. During training, the network adjusts to reduce the error on the training patterns. The accuracy of training is found out by the parameters such as NN architecture, activation function, inputs, number of hidden neurons in hidden layer, and updating of weights. The sigmoidal activation function is used in the hidden layer and softmax activation function used in the output layer. The number of hidden neurons are chosen using trial rule and weights are updated by means of SCG algorithm. In this work the features presented in previous section such as jitter, shimmer, HNR, GNE, SNR, MFCC and RASTA-PLP extracted from normal, ASD and DS voices are given as input to the FFNN for classification.

4 Results and Discussion

Experiments are conducted for different normal and pathological datasets. All the speech samples are the phonation of sustained vowel /ah/. Perturbation measures such as jitter and shimmer, noise parameters such as HNR, GNE and SNR, Cepstral features like MFFC and RASTA-PLP feature set are extracted from each voice sample and fed to a Feed-Forward Neural Network (FFNN) trained with Scaled Conjugate Gradient (SCG) algorithm to differentiate their classes. Then the performance analysis is done and further compared with other classification methods.

4.1 Database

The database used in this study is constructed with 79 Autism Spectrum Disorder (ASD) samples, 77 Down Syndrome (DS) and 99 normal samples. All the speech samples are the phonation of sustained vowel /ah/ and are collected from six ASD children (three boys, three girls), six DS children (two boys, four girls) and seven controls (three boys, four girls). All children are aged between 5–14 years. The acoustic samples are recorded in noise free location using Audio Editor 2016 software installed in Laptop. speech samples are wave files, in PCM format and in mono mode at a sampling rate of 24,000 kHz and 16-bit resolution. The speech signals are preprocessed using silence removal and windowing techniques. The size of hamming window used is 30 ms in length.

4.2 Performance Evaluation

The relevant features are extracted from the preprocessed voice samples. The features jitter, shimmer, HNR, GNE, SNR, MFCC, RASTA-PLP are computed for each normal and pathological voice sample. Four distinguished features such as mean, standard deviation, skewness, kurtosis are computed for HNR and GNE extracted across each analysis window and they form the feature vector. For extracting MFCC features, windowing with 30 ms length and 15 ms time shift are used. From each frame, 12 static MFCC features, log energy, the 0^{th} cepstral coefficient and, two dynamic features delta and delta-delta are computed. Mean and standard deviation of the 16 features computed from each analysis frame forms the feature vector. Moreover, 11^{th} order all-pole modeling is done to extract 12 RASTA-PLP features from hamming windowed speech frames of 25 ms length with 10 ms overlap. Mean and standard deviation are calculated for RASTA-PLP features computed from each frame and they form the feature vector. The performance of the three proposed systems is evaluated for each feature and the FFNN classifier.

In all the three experiments the 70% of the data are used for training, 15% for cross-validation and 15% for testing. A six-fold cross-validation is used. Both the individual and the combined feature vectors constitute the nodes of the input layer. The network consists of one hidden layer having *sigmoidal* activation function. The number of neurons in the hidden layer are chosen by trail rule. The number of neurons in the output layer corresponds to number of classes. The optimal learning rate is 0.05, the optimal momentum is 0.3. Further 2000 training epochs are needed for achieving the lowest Mean Square Error (MSE), i.e., MSE = 0.0001. Here, classes correspond to normal, Automatic Spectrum Disorder (ASD) and Down Syndrome (DS). The *softmax* activation function is used in the output layer. Experiments are also done to see if the combination of features could improve the performance of the proposed models. The performance parameters of the FFNN classifier are calculated using the following relationships:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} x100 \tag{26}$$

$$Sensitivity = \frac{TP}{TP + FN} x100 \tag{27}$$

$$Specificity = \frac{TN}{TN + FP} x100 \tag{28}$$

where,

TP (True Positive) = Number of pathological samples classified as pathological.

TN (True Negative) = Number of normal samples classified as normal.

FP (False Positive) = Number of normal samples classified as pathological.

FN (False Negative) = Number of pathological samples classified as normal.

In ASD detection and DS detection, the autistic samples and DS samples respectively form the positive class and the normal samples from the negative class. In ASD-DS classification the autistic samples form the positive class and DS samples form the negative class. In addition to accuracy, sensitivity and specificity, the Area Under Receiver Operating Characteristics (AUC) is also computed to show the result in more compact form [35]. Receiver Operating Characteristic (ROC) is also a popular tool in medical decision making [36]. It reveals diagnostic accuracy expressed in terms of sensitivity and specificity. The AUC is a single scalar representing an estimation of the expected performance of the system [37,38].

4.3 Individual/Combined Features and FFNN

The performance of the classifier is analyzed by first considering the individual feature sets such as jitter, shimmer, GNE, HNE, SNR, MFCC and RASTA-PLP. The corresponding results are shown in Tab. 1.

Table 1: Results obtained for DS detection/ASD-DS classification with individual features and FFNN

ASD detection with individual features and FFNN					
Individual features	No. of hidden layer neurons	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Jitter	17	88.88	100	83	0.972
Shimmer	9	83.33	85.71	81.81	0.870
HNR	13	77.8	57.14	81.81	0.766
GNE	8	83.33	100	66.66	0.839
SNR	20	89.4	90.2	85.6	0.813
MFCC	**6**	**100**	**100**	**100**	**1**
RASTA-PLP	**11**	**100**	**100**	**100**	**1**
DS detection with individual features and FFNN					
Jitter	9	81.25	100	66.66	0.709
Shimmer	10	81.25	80	81.81	0.945
HNR	9	94.73	100	91.66	1
GNE	17	89.47	100	89	0.904
SNR	20	92.4	86.2	83.1	0.845
MFCC	**7**	**100**	**100**	**100**	**1**
RASTA-PLP	**15**	**100**	**100**	**100**	**1**

(Continued)

Table 1 (continued).

ASD-DS classification with individual features and FFNN					
Jitter	8	81.25	90	66	0.809
Shimmer	10	81.25	80	81.81	0.761
HNR	18	100	100	100	1
GNE	5	73.68	63.63	87.5	0.920
SNR	20	90.4	79.2	79.5	0.834
MFCC	**9**	**100**	**100**	**100**	**1**
RASTA-PLP	**14**	**85.7**	**96.66**	**85.7**	**0.95**

It is observed that, the MFCC and RASTA-PLP features provide 100% accuracy for ASD detection with 6 and 11 hidden neurons respectively. The maximum accuracy obtained for jitter is 88.88% with 8 hidden layer neurons; for shimmer, 83.33% with 9 hidden neurons; for GNE, 83.33% with 8 hidden neurons; and for HNR, 77.8% with 13 hidden neurons. The AUC for ASD detection is 1 for MFCC and RASTA-PLP and 0.972, 0.870, 0.839 and 0.766 for jitter, shimmer, GNE, and HNR respectively. It is also seen that, the classification accuracy obtained for DS detection is 100% with MFCC and RASTA-PLP for 7 and 15 hidden neurons, respectively. The accuracy of 94.73% is attained with HNR; 89.47% with GNE; and 81.25% with jitter and shimmer. For DS identification, the AUC is 1 for MFCC, RASTA-PLP, and HNR. The AUC for jitter, shimmer, GNE is 0.709, 0.945 and 0.904, respectively. Similarly, the HNR and MFCC can provide 100% classification accuracy with 18 and 9 hidden neurons respectively for ASD-DS classification. The accuracy attained for classification between AD as DS is 81.25%, 81.25% and 73.68% for jitter, shimmer, and GNE, respectively. The AUC for classification between AD and DS is 1 for both MFCC and HNR features. The jitter, shimmer, GNE, and RASTA-PLP can provide 0.809, 0.761, 0.920 and 0.797 AUC, respectively.

The MFCC and RASTA-PLP provide 100% Accuracy for ASD identification. This is because MFCC and RASTA-PLP are Fourier transform-based features whereas jitter, shimmer, HNR, and GNE are time-domain features. It is also evident that for ASD-DS classification, MFCC provides 100% Accuracy. This result is consistent with previous studies that, spectral and cepstral based features are best suited for discriminating voice with hoarseness from normal voice [39].

The performance of the FFNN classifier is also analyzed by using combined feature sets. That is, using combination of features such as jitter and shimmer, HNR and GNE, HNR and SNR, MFCC and RASTA-PLP. The performance measures obtained are shown in Tab. 2.

From Tabs. 1 and 2, it is found that, the combined feature set provides less performance when compared to the individual features, for most of the cases. It is also noted that the frequency domain based feature sets, i.e., MFCC and RASTA-PLP with the FFNN classifier produce improved performance measures.

4.4 Processing Time

The experiments are carried out using *i5* core pendium processor and implemented in MATLAB software. The computation or processing time for both detection and classification for FFNN using features are shown in Fig. 5.

It is evident that, the MFCC and the RASTA-PLP feature sets produced less processing time, when compared to the other feature sets. The method of finding jitter and shimmer depends on the fundamental

frequency. It is very difficult to extract the fundamental frequency and it needs high processing time as shown in Fig. 5. For the computation of HNE, the process of finding autocorrelation is needed. Thus the proceeding time is high as shown in Fig. 5. Similarly, for the calculation of SNR depends on the nature of the noise present in the signal. As discussed in Section 2.2.2, the computation of GNE requires the necessity of extracting Hilbert envelopes for each glottal cycle and finding the cross correlation of the envelope pairs and obtain the maximum value of the envelope pair. This procedure requires higher processing time as shown in Fig. 5. But the proceesing time needed for the computation of MFCC and RASTA-PLP is less. These features can be calculated easily using the inbuilt routines present in the MATLAB.

Table 2: Results for ASD/DS detection and ADS-DS classification with combined features and FFNN

Combined feature sets	No. of hidden layer neurons	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
ASD detection with combined features					
Jitter and shimmer	8	88.9	100	83.3	0.902
HNR and GNE	15	77.8	80.0	80.0	0.837
HNR and SNR	7	53.2	79.2	85.2	0.85
MFCC and RASTA-PLP	10	80.7	82.6	82.7	0.89
DS detection with combined features					
Jitter and shimmer	11	93.75	100	90.9	0.981
HNR and GNE	15	77.8	80.0	80.0	0.837
HNR and SNR	12	86.3	93.2	67.4	0.851
MFCC and RASTA-PLP	11	78.3	83.6	87.4	0.862
ADS-DS classification with combined features					
Jitter and shimmer	17	81.3	100	40	0.927
HNR and GNE	4	100	100	100	1
HNR and SNR	12	79.5	84.2	92.6	0.925
MFCC and RASTA-PLP	13	80.7	82.6	82.7	0.876

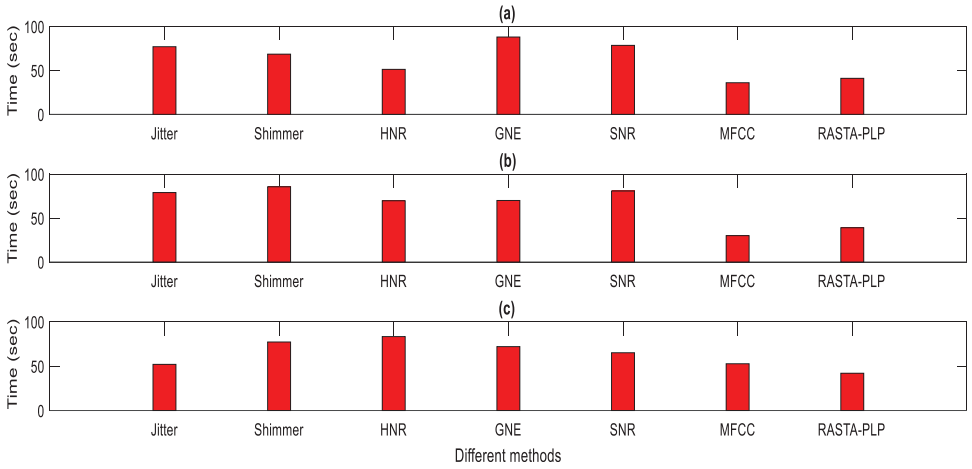

Figure 5: Processing time for both detection and classification for FFNN using features

4.5 Effect of Neurons in the Hidden Layer

The number of neurons that should be retained in the hidden layer need to be calculated. If the number of neurons are very less, "Underfitting" may occur. Moreover, if more neurons are present in the network, then "Overfitting" may occur. Several methods are used till now which do not provide the exact formula for calculating the number of hidden layer as well as number of neurons in each hidden layer. In this work, the number of hidden neurons are fixed by means of trial rule.

Firstly, few random number of neurons are fixed and samples are allowed to train on it. If the network does not converge after a sensible period of time, repeat the training process, so that it is assured that it has not dropped into local minima. If the network still does not converge, few more neurons are added in the layer and allow it to train until the network converges and produces high accuracy. Fig. 6 show the variation of accuracy in percentage for different number of hidden neurons with the individual/ combined features and FFNN.

It is clear that, the jitter shows maximum classification rate, if the number hidden neurons is 17, as shown in Fig. 6a: 1. The MFCC shows the classification rate of 100%, if there are 6 hidden neurons. Likewise, the number of hidden neurons of the FFNN classifier for all the features are chosen. Fig. 6a: 2 shows the variation of hidden layer neurons with respect to accuracy for ADS detection using combined features and FFNN.

In the same way, Figs. 6b and 6c show the variation of accuracy of the FFNN classifier with respect to the number of hidden layer neurons for DS detection and ADS-DS classification for the individual and combined features, respectively.

The number of hidden neurons are randomly varied in every step. Finally choose the hidden neurons in such a way that the accuracy is maximum and the MSE attained is very less.

4.6 Comparison with SVM and Navie Bayes Classifier

The performance of the FFNN classifier is also analyzed by comparing the results with two more classifiers namely SVM and Navie Bayes estimation classifier with the same set of features. In SVM approach, a hyperplane or set of hyperplanes are constructed in a high or infinite dimensional space, which can be used for classification. Spontaneously, a good separation is accomplished by the hyperplane that has the largest distance to the nearest training data point of any class. The LIBSVM is trained on the relevant feature vectors using RBF kernel function. The LIBSVM is used to test these feature vectors. The investigation is carried out by varying cost values for the RBF kernel. The Naive Bayes is a probabilistic classifier in which, for a document d, out of all classes $c \in C$ the classifier returns a new class that has the maximum posterior probability. Tab. 3 shows the comparison results of FFNN classifier with SVM and Naive Bayes classifiers using MFCC and RASTA-PLP features.

The Navie Bayes classifier achieved the accuracy of 83.3% with the sensitivity 90.7%, specificity 81.9%, AUC 0.79 using MFCC features and 72.7% with the specificity of 84.7%, specificity 79.3%, AUC 0.83 using RASTA-PLP features for ASD-DS classification. These values are found to be less, when compared to the results achieved using FFNN classifier. It is clearly noticed that, for all the cases, the FFNN classifier provides promising results. It is also observed that, very good performance is attained for ASD and DS detection using FFNN classifier.

From the analysis of Tabs. 3 and 4, it is evident that, the performance of the individual feature set, i.e., MFCC and RASTA-PLP features combined with FFNN classifier gives significant improvement in terms of accuracy, sensitivity, specificity and AUC for both the detection and classification of Down Syndrome (DS) and Automatic Spectrum Disorder (ASD). It is also concluded that, the MFCC features and FFNN produced highest performance measures.

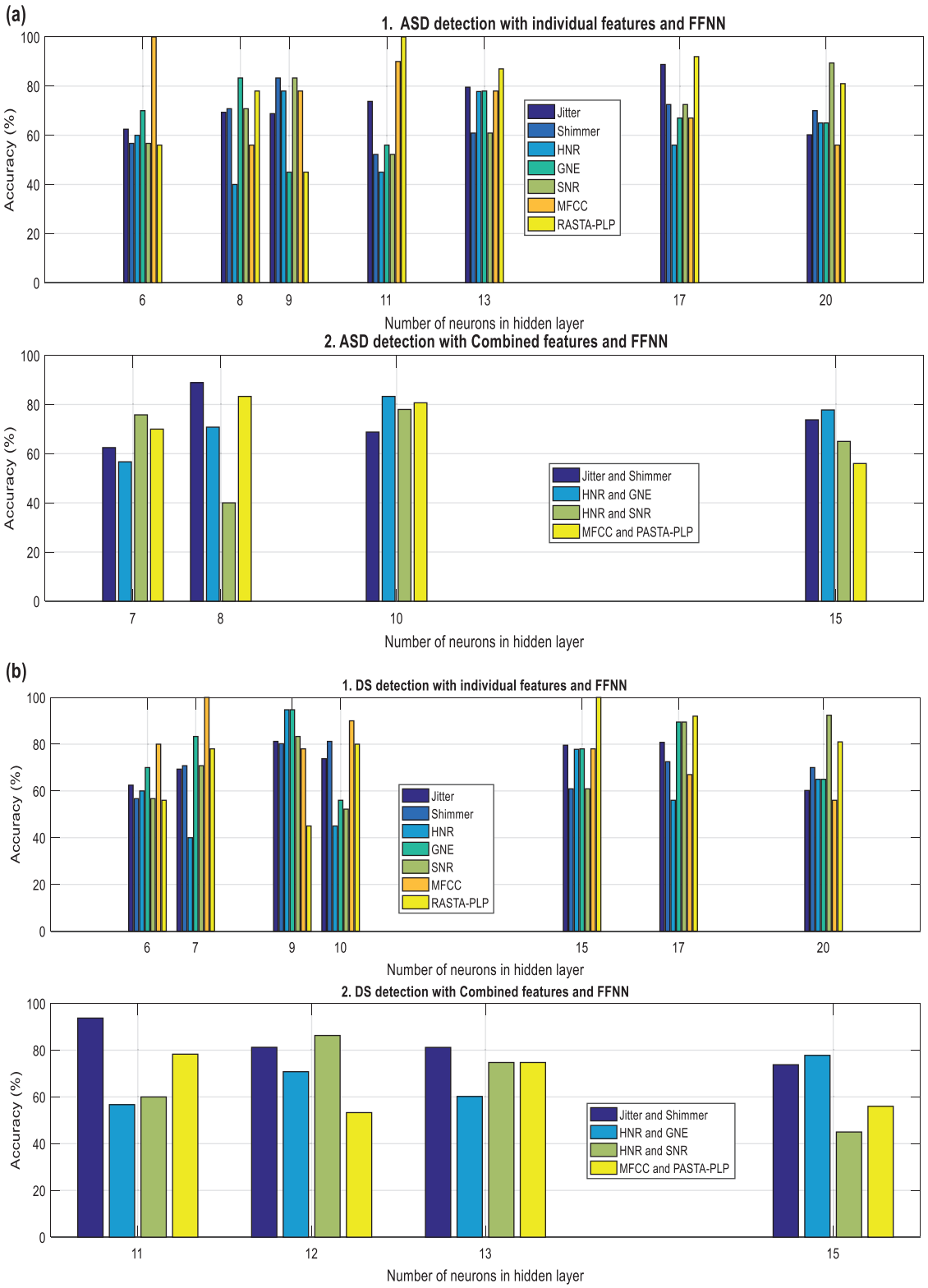

Figure 6: (continued)

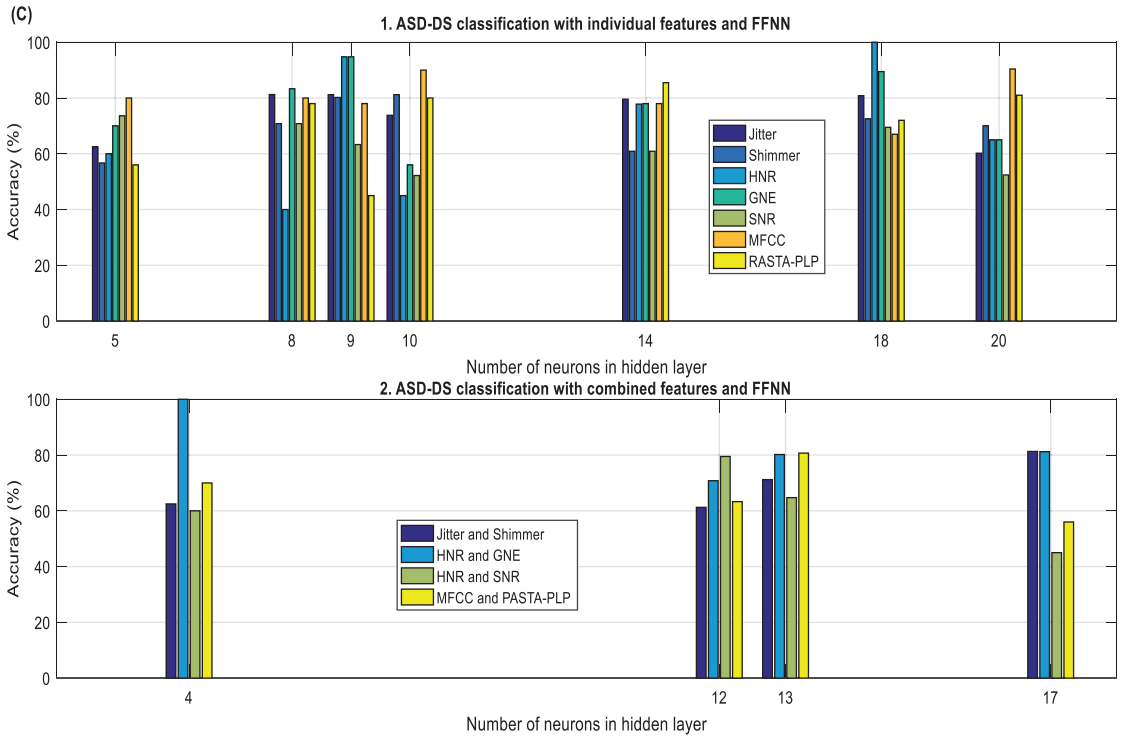

Figure 6: (a) Accuracy with variation in the hidden neurons for ASD detection (b) Variation of classification rate with the hidden neurons for DS detection (c) Variation of accuracy with the hidden neurons for ADS-DS detection

Table 3: Comparison results of FFNN classifier with SVM and Naïve Bayes classifiers using MFCC and RASTA-PLP features

Performance measures	MFCC			RASTA-PLP		
	SVM	Naive Bayes	FFNN	SVM	Naive Bayes	FFNN
ASD detection						
Accuracy	78.5	91.2	**100**	78.5	91.2	**100**
Sensitivity	89.7	82.4	**100**	89.7	82.4	**100**
Specificity	75.7	79.3	**100**	89.3	87.9	**100**
AUC	0.89	0.78	1	0.93	0.87	**1**
DS detection						
Accuracy	85.6	89.7	**100**	90.8	79.2	**100**
Sensitivity	78.5	91.2	**100**	92.4	83.2	**100**
Specificity	89.7	82.4	**100**	89.3	87.9	**100**
AUC	0.86	0.82	**1**	0.92	91.2	**1**
ASD-DS classification						
Accuracy	77.2	83.3	**100**	85.1	72.7	**85.7**
Sensitivity	90.7	86.3	**100**	88.4	84.7	**95.66**
Specificity	81.9	78.9	**100**	75.7	79.3	**85.7**
AUC	0.79	0.76	**1**	74.3	0.83	**0.95**

Table 4: Comparison results of FFNN classifier with SVM and Naïve Bayes classifiers using combined features

Performance measures	Jitter and Shimmer			HNR and GNE			HNR and SNR			MFCC and RASTA-PLP		
	SVM	Naive Bayes	FFNN	SVM	Naive Bayes	FFNN	SVM	Naive Bayes	FFNN	SVM	Naive Bayes	FFNN
ASD detection												
Accuracy	75.3	69.2	**88.9**	70.8	80.5	**77.8**	84.2	70.7	**53.2**	79.6	81.2	**80.7**
Sensitivity	81.2	68.3	**100**	83.6	83.7	**80.0**	72.6	71.6	**79.2**	74.8	79.4	**82.6**
Specificity	79.4	75.2	**83.3**	78.4	81.2	**80.0**	75.8	83.4	**85.2**	82.5	0.82	**82.7**
AUC	0.82	79.4	**0.90**	0.80	78.6	**0.83**	0.80	0.78	**0.85**	0.82	0.80	**0.89**
DS detection												
Accuracy	79.6	70.7	**93.7**	84.2	69.2	**77.8**	81.2	76.9	**86.3**	80.5	70.8	**78.3**
Sensitivity	74.8	71.6	**100**	72.6	68.3	**80.0**	79.4	72.8	**93.2**	83.7	83.6	**83.6**
Specificity	82.5	83.4	**90.9**	75.8	75.2	**80.0**	0.82	70.4	**67.4**	81.2	78.4	**87.4**
AUC	0.82	0.78	**0.98**	0.80	79.4	**0.83**	0.74	0.76	**0.85**	78.6	0.79	**0.86**
ASD-DS classification												
Accuracy	76.9	80.2	**81.3**	70.8	81.2	**100**	79.6	80.5	**79.5**	69.2	70.7	**80.7**
Sensitivity	72.8	86.9	**100**	83.6	79.4	**100**	74.8	83.7	**84.2**	68.3	71.6	**82.6**
Specificity	70.4	79.0	**82.4**	78.4	0.82	**100**	82.5	81.2	**92.6**	75.2	83.4	**82.7**
AUC	0.76	0.87	**0.92**	0.92	0.89	**1**	0.82	78.6	**0.92**	79.4	0.78	**0.87**

The results of the experiment indicate the existence of atypical voice patterns in ASD and DS speech, so that their detection and classification are possible with 100% accuracy. Thus it is concluded that, in ASD and DS identification, higher performance was achieved with MFCC and RASATA PLP features, while for discriminating ADS from DS a maximum accuracy of 100% was achieved with features HNR and MFCC. This system provides a firm foundation for automatic detection of ASD and DS using acoustic analysis of speech.

5 Conclusion

In this work the capability of the three sets of features and the ANN classifier in performing the three proposed tasks (i) detecting ASD (ii) detecting DS (iii) ADS-DS classification have been discussed. Advanced signal processing methods are presented to derive the important attributes hidden in the voice signals. Three different set of features are discussed in detail. From experimental results, we can conclude that the cepstral features MFCC and RASTA-PLP are best suited for both detection and classification of voice pathological signals. It is alo also found that the combination of MFCC features and FFNN gives promising results, when compared to other set of feature and other classifiers.

Acknowledgement: The authors are extremely grateful to Dr. Charles Russel, Managing Director of Rejoice Autism School, Nagercoil and Mrs. Jessila Banu, Headmistress, Nanjil Oasis Happy Centre for Mentally Retarded Children, Nagercoil for granting permission and helping us to collect autism and down syndrome voice samples that are used in this work. The authors would like to thank Mrs. T. Devi, Principal, Devi Matriculation School, Thuckalay for granting permission to collect Normal voice samples used in this research.

Funding Statement: The author(s) received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Islam, R., Tarique, M., Abdel-Raheem, E. (2020). A survey on signal processing based pathological voice detection techniques. *IEEE Access, 8,* 66749–66776. DOI 10.1109/ACCESS.2020.2985280.
2. Rudra, A., Belmonte, M. K., Soni, P. K., Banerjee, S., Mukerji, S. (2017). Prevalence of autism spectrum disorder and autistic symptoms in a school-base cohort of children in Kolkata, India. *Autism Research, 10(10),* 1597–1605. DOI 10.1002/aur.1812.
3. Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child, (2),* 217–250.
4. Lord, C., Rutter, M., Le Couteur, A. (1994). Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 24(5),* 659–685. DOI 10.1007/BF02172145.
5. Lord, C., Risi, S. (2000). The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders, 30(3),* 205–223. DOI 10.1023/A:1005592401947.
6. Kumar, M. P., Sharma, R. K. (2001). Estimation ans statistical analysis of human voice parameters to investigate the influence of Psycological stress and to determine the vocal tract trtansfer function of an individual. *Journal of Computer networks and Communications,* 1–17.
7. Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., Cluibley, E. (2001). The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders, 31(1),* 5–17. DOI 10.1023/A:1005653411471.
8. Albertini, G., Bonassi, S., Dall'Armi, V., Giachetti, I., Giaquinti, S. et al. (2010). Spectral analysis of the voice in Down Syndrome. *Research in Developmental Disabilities, 31(5),* 995–1001. DOI 10.1016/j.ridd.2010.04.024.
9. Lee, M. T., Thorpe, J., Verhoeven, J. (2009). Intonation and phonation in young adults with Down Syndrome. *Journal of Voice, 23(1),* 82–87. DOI 10.1016/j.jvoice.2007.04.006.
10. Moura, C. P., Cunha, L. M., Vilarinho, H. (2008). Voice parameters in children with Down Syndrome. *Journal of Voice, 22(1),* 34–42. DOI 10.1016/j.jvoice.2006.08.011.
11. Jeffery, T., Cunningham, S., Whiteside, S. P. (2018). Analysis of sustained vowels in Down Syndrome (DS): A case study using Spectrograms and Perturbation data to investigate voice quality in four adults with DS. *Journal of Voice, 32(5),* 644.e11–644.e24. DOI 10.1016/j.jvoice.2017.08.004.
12. Eybe, F., Weninger, F., Grob, F., Schuller, B. (2013). Recent developments in openSMILE, the munich open-source multimedia feature extractor. *Proceedings of the 21st ACM International Conference on Multimedia,* pp. 835–838.
13. Sampaio, M., Lucia, M., Masson, V. (2020). Effects of fundamental frequency, vocal intensity, sample duration and vowel context in Cepstral and spectral measures of Dysphonic voices. *Journal of Speech, Language and Hearing Research, 63,* 120–127.
14. Deng, J., Cummins, N., Schmitt, M., Quian, K. et al. (2017). Speech-based diagnosis of autism spectrum condition by generative adversial network representations. *Proceedings of the ACM Digital Health,* pp. 53–57. London.
15. Seyyed Hamid, Ebrahimi Motlagh, Hadi Moradi, Hamidreza Pouretemad (2013). Using general sound descriptors for early autism detection. *Control Conference (ASCC),* pp. 125–130.
16. Sundarsana, K., Alku, P. (2020). Analysis and detection of pathological voices using glottal source features. *IEEE Journal of Selected Topics in Signal Processing, 14,* 57–63.
17. Berument, S. K., Rutter, M., Lord, C., Pickles, A., Bailey, A. (2000). Autism screening questionnaire: Diagnostic validity. *Psychiatry-Interpersonal and Biological Processes, 175,* 444–451.
18. Gorlin, R., Cohen, M. M., Levin, L. S. (1990). Chromosomal syndromes: Common and/or well-known syndromes: Trisomy 21 Syndrome (Down Syndrome). *Syndromes of the Head and Neck, (3),* 33–40.
19. Yoram, S., Bonneh (2011). Abnormal speech spectrum an increased pitch variability in young autistic children. *Frontiers in Human Neuroscience, 4,* 237–242.

20. Dankovicova, Z., Sovak, D., Drotar, P., Vokorokos, L. (2018). Machine learning approach for dysphonia detection. *Applied Sciences,* 1–12.
21. Sun, X. (2002). Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. *IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. 126–134.
22. Tsanas, A. (2012). *Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning* (*D Phil thesis)*. University of Oxford, Oxford, UK.
23. Anastasis., Kounoudes., Patrick A. Naylor., Mike Brookes (2002). The DYPSA algorithm for estimation of glottal closure instants in voiced speech. *IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. 349–352.
24. Kim, K. S., Seo, J. H., Song, C. G. (2010). An acoustical evaluation of knee sound for non-invasive screening and early detection of articular pathology. *Journal of Medical Systems, 36(2),* 715–722. DOI 10.1007/s10916-010-9539-3.
25. Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proceedings, 17,* 97–110.
26. Michaelis, D., Gramss, T., Strube, H. W. (1997). Glottal-to-noise excitation ratio-a new measure for describing pathological voices. *Acta Acustica United with Acustica, 83(4),* 700–706.
27. He, L., Zhang, J. (2015). Automatic evaluation of hypernasality based on a cleft palate speech database. *Journal of Medical Systems, 39(5),* 242.
28. Huang, X. D., Acero, A., Hsiad., Ho, H. W. (2001). *Spoken language processing: A guide to theory, algorithm, and system development,* pp. 316–318. Prentice Hall.
29. Hermansky, H., Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing, 2(4),* 578–589. DOI 10.1109/89.326616.
30. Ali, Z. (2016). Automatic voice pathology detection with running speech by using estimation of auditory spectrum and cepstral coefficients based on the All-Pole model. *Journal of Voice, 30(6),* 757–761. DOI 10.1016/j.jvoice.2015.08.010.
31. Daqrouq, K., Tutunji, T. A. (2015). Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers. *Applied Soft Computing, 27(1–3),* 231–239. DOI 10.1016/j.asoc.2014.11.016.
32. Møller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks, 6(4),* 525–533. DOI 10.1016/S0893-6080(05)80056-5.
33. Jose Orozco, G., Carlos, A., Garcia, R., Erro, L. E. (2003). Detecting pathologies from infant cry applying scaled conjugate gradient neural networks. *Proceedings of European Symposium on Artificial Neural Networks, d-side public*, pp. 349–354.
34. Haseena, H. H., Mathew, A. T., Paul, J. K. (2011). Fuzzy clustered probabilistic and multi layered feed forward neural networks for electrocardiogram Arrhythmia classification. *Journal of Medical Systems, 35(2),* 179–188. DOI 10.1007/s10916-009-9355-9.
35. Nicolas Sáenz-Lechon, N., Godino-Llorente, J. I., Osma-Ruiz, V. (2006). Methodological issues in the development of automatic systems for voice pathology detection. *Biomedical Signal Processing and Control, 1(2),* 120–128. DOI 10.1016/j.bspc.2006.06.003.
36. Hanley, J. A., McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143(1),* 29–36. DOI 10.1148/radiology.143.1.7063747.
37. Michaelis, D., Fröhlich, M., Strube, H. W. (1998). Selection and combination of Acoustic features for the description of pathologic. *Journal of the Acoustical Society of America,* 121–129.
38. Cappe, O., Moulines, E. (2009). *Erik Moulines, Inference in Hidden Markov Models*. Springer-Verlag, 172–180.
39. Orozco-Arroyave, J. R. (2015). Characterization methods for the detection of multiple voice disorders, neurological, functional, and laryngeal diseases. *IEEE Journal of Biomedical and Health Informatics, 19(6),* 1820–1828. DOI 10.1109/JBHI.2015.2467375.