A REVIEW ON CLASSIFICATION OF LARYNGEAL PATHOLOGY USING AUTOMATED SPEECH ANALYSIS

[1]Jarecha Dipa, [2]Prof.Mitul Patel
[1]M.E. Student, [2]Professor
[1]Department of Biomedical Engineering,
[1]Government Engineering College, Gandhinagar, India

Abstract: Speech disorders are a common physical problem that can be encountered today and can cause serious problems in the long term. Due to the irregular vocal fold vibration, brain damage, hearing loss, stroke, speech disorders occur. Nowadays Speech disorders are diagnosed using laryngoscope and endoscopy, making the processing time consuming, painful, discomfort and very costly to the patient which is invasive procedure, so it is needed to allocate a non-invasive mechanism. This paper gives a review on different techniques used for automated detection of laryngeal pathologies. This acoustic voice analysis techniques helps in development of low-cost screening tool for detection of laryngeal pathologies.

Key-words – Laryngeal Pathologies, Feature Extraction, Classification Algorithm

I. INTRODUCTION

Speech is very important to our daily life for communication with each other. Speech disorders disturb the process of a person to produce a speech signal to form words or sentences. Speech disorders can cause frequent hoarseness on the voice in the general population. Due to muscle weakness, vocal cord damage, brain injury, vocal cord nodules, respiratory weakness, vocal cord polyps, vocal cord paralysis, hearing impairment speech disorders occur. Nowadays speech pathology is a diagnosis by the invasive methods endoscopy or medical micro laryngoscopy. Which distress the individual, painful, discomfort and expensive also. These methods are time-consuming and some advanced pieces of equipment availability are limited in some rural areas. Which are invasive methods so it is required to provide them non-invasive methods.

Speech and voice diseases cause acoustic changes in voice signal. Therefore, Voice signal can be used for diagnosing them. Based on extracting its acoustic features, they are classified as normal and abnormal speech using classification algorithm. Artificial Intelligence takes place in the medical field for providing easily adopted techniques for diagnosis. Voice signal-based analysis enormous potential for research and developing new analytical tool, in which software based non-invasive methods are used for analysis. It provides an essential diagnostic device to medical experts, common persons for diagnosis of laryngeal pathologies essentially in primary stage.

Objective of this paper is to provide a review on different non-invasive methods used for speech disorder diagnosis and to compare various online data base available, various features used for classification, methodology and classification techniques used by researchers for classification of laryngeal pathologies.

Organization of this document is as follows. Section II describes literature review. Section III gives information about materials and methods. Section IV gives conclusion and section V draws conclusion.

Linguist and speech pathologists have diagnosed characteristic irregularities in voice patterns of patients with certain medical conditions like Asthma, Depression, Autism Spectrum Disorders, Parkinson's disease (PD), Alzheimer's disease, Schizophrenia, Larynx cancer, etc. Scientists as well as medical researchers are exploring the correlation to quantify these variations in voice corresponding to various medical conditions for designing the voice analysis systems for diagnosis and treatment of respective health disorders in a controlled clinical environment [1]. These voice analysis methods facilitate clinicians in screening patients and scaling the progress of ongoing treatment. Numerous research studies have been conducted for classifying these acoustic [2], prosodic [3], emotional [4] or lexical voice features for extracting health information of the subjects.

Table 2.1 Literature review

Sr.No.	Title	Author	Year	Publication	Description
1.	**Acoustic Analysis of Vocal Dysphonia [10]**	**João Paulo Teixeira Paula Odete Fernandesa**	**2015**	**Conference on enterprise Information Systems / International Conference on Project management / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2015 October 7-9, 2015**	In this paper, they have done the statistical analysis of a set of voice parameters **jitter, shimmer and HNR.** Database collected for **dysphonia, functional dysphonia, hyper functional dysphonia, psychogenic dysphonia** and healthy person from the online database Saarbrücken Voice Database (SDB). Total **154** subject's database is collected during phonation of vowel /a/, /i/, /u/ for high, low and normal tone. They have done the statistical analysis on the basis of mean and standard deviation and no separate classification is made for **men and women.** Moreover, they have done only classification between normal and pathological voice.
2.	**Acoustic Analysis for Detection of Voice Disorders Using Adaptive Features and Classifiers [11]**	**Mohamed FEZARI Fethi AMARA Ibrahim M. M. El-EMARY**	**2014**	**International Conference on Circuits, Systems and Control**	In this paper, they have the methods of acoustic voice analysis (AVA) of non-neurological voice disorder such as **chronical laryngitis** and **Vocal fold nodules.** Data is collected from the **online database Saarbrücken Voice Database (SDB) for 81 patients.** **Mel frequency cepstral coefficients MFCC, Jitter & shimmer** following features are extracted for analysis purpose. **Gaussian mixture model (GMM)** is used for classification purpose and they got the accuracy of **82%** using this method.
3.	**Recognition of Dysarthric Speech Using Voice Parameters for Speaker Adaptation**	**Chitralekha Bhat Bhavik Vachhani Sunil Kopparapu**	**2016**	**International speech communication association**	In this paper, they have collected data from **Universal Access (UA) speech corpus for 28 patients.**

	and Multi-taper Spectral Estimation [5]				**The recording material consisted of 455distinct words with 10 digits, 26international radio alphabets,19 computer commands, 100 common words and 300 uncommon words that were distributed into three blocks. Three blocks of data were collected for each speaker such that in each block speaker recorded the digits, radio alphabets, computer commands, common words and 100 of the uncommon words. Thuseach speaker recorded 765 isolated words.** After collecting data following features are extracted: **jitter, shimmer, F0 and Noise Harmonic Ratio (NHR) and multi taper spectral estimation for counting** Mel Frequency Cepstral Coefficients (MFCC). After that they have used hidden Markov model (HMM) recognition system and a Gaussian Mixture Model (GMM) for dysarthric speech recognition. They have analyzed that hidden Markov model (HMM) recognition system is fared better than a Gaussian Mixture Model (GMM).
4.	**Voice Disorder Identification by Using Machine Learning Techniques [6]**	**LAURA VERDE GIUSEPPE DE PIETRO GIOVANNA SANNINO**	**2018**	**IEEE**	In this paper, Authors have selected a subset of voice samples from the ``Saarbrucken Voice Database'' (SVD). SVD database is a collection of **2041** voice recordings, containing voices from healthy and pathological individuals, published online by the Institute of Phonetics of the University of Saarland. From this database they have selected a total of 1370 samples. **They have used features like** Fundamental

					Frequency, Jittler, Shimmer, HNR, MFCC and First and second derivatives of cepstral coefficient**.** For classification they have used **machine learning classifier like Support vector machine. They have classified normal and pathological voice using this method. They got accuracy of 71% using this method**
5.	**Voice Disorders Identification Using Multilayer Neural Network [7]**	**Lotfi Salhi TalbiMourad Adnene Cherif**	**2010**	**The International Arab Journal of Information Technology,2010**	In this paper, authors have conducted a research in a supervised mode for classification between normal and vocal pathologies. They have analyzed 2 features **pitch and formants** and done classification using multilayer neural network.by using this method they got accuracy of **80%** on **30** patients
6.	**Classification of Dysphonic Voice: Acoustic and Auditory-Perceptual Measures [8]**	**Eadie Tanya Doyle Philip**	**2002**	**Journal of voice: official journal of the Voice Foundation**	In this paper authors have developed a system for detecting **dysphonia**. They have used non parametric measures such as **long-term average spectral measures** and **glottal noise measures** as a and on linear prediction modelling, which in turn formed the inputs into conditional logistic regression analysis**. 100 percent** of accuracy in voice pathology detection was achieved using this method but This research included only **24 patients.**
7.	**ARTIFICIAL NEURAL NETWORK BASED PATHOLOGICAL VOICE CLASSIFICATION USING MFCC FEATURES[9]**	**V. Srinivasan. Ramalingam P. Arulmozhi**	**2014**	**International Journal of Science, Environment ISSN 2278-3687 (O)and Technology,2014**	In this paper, they have analyzed pathological voices with the aid of the speech signals recorded from the patients.**20 persons** are selected having pathological voice quality for this analysis. They have extracted only one feature Mel-Frequency Cepstral Coefficients (MFCC) feature from audio recordings for analysis. They have used method for the identification and classification of pathological voice are

					artificial Neural Network. Multilayer Perceptron Neural Network (MLPNN)Generalized Regression Neural Network (GRNN) and Probabilistic Neural Network (PNN) .100% percent accuracy was achieved using this method.

III. Materials and Methods:

General Block-diagram of Existing System:

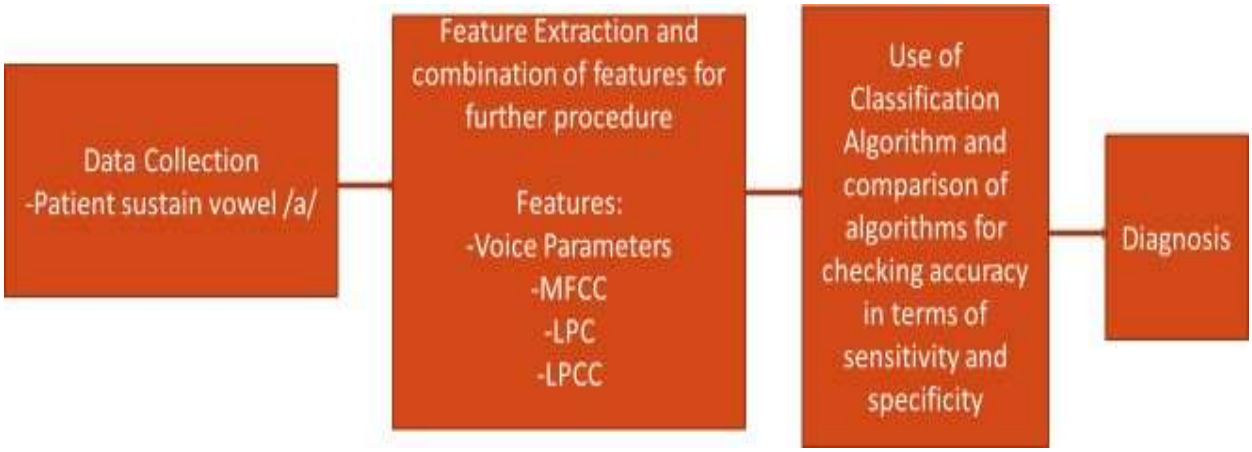

Figure 3.1 General Block- Diagram of the system

Figure 3.1 shows the functional diagram of the proposed methods. In this proposed block diagram first step is data collection in which the voice data is collected through the microphone form normal and abnormal persons. After that, this speech signal's feature is extracted using a different algorithm. these features are then fed to the classifier for classification of the normal and abnormal pathological signal.

3.1 Database

For recording of voice database of abnormal patient generally, they use microphone, mobile phone or laptop. This hardware also plays an important role in data analysis. Patient sustained vowel sound a, i, u or some other words. Recorded file can be stored in different format such as wav, mp3 and PCM format having different fundamental frequency such as 44100hz,8000hz,16000hz etc. Online database is also available that can be used for analysis. Online database includes Saarbrücken voice database machine learning repository database, MEEI voice and speech database is used.

3.2 Pre-processing

Pre-emphasis - To enhance the accuracy and efficiency of the extraction processes, speech signals are normally pre-processed before features are extracted. The aim of this process is to boost the amount of energy in the high frequencies. The drop in energy across frequencies (which is called spectral tilt) is caused by the nature of the glottal pulse. Boosting the high frequency energy makes information from these higher formants available to the acoustic model. The pre-emphasis filter is applied on the input signal before windowing.

Framing - It is a process of segmenting the speech samples obtained from the analog to digital conversion (ADC), into the small frames with the time length within the range of 20-40 ms. Framing enables the non-stationary speech signal to be segmented into quasi-stationary frames, and enables Fourier Transformation of the speech signal. It is because, speech signal is known to exhibit quasi-stationary behavior within the short time period of 20-40 ms. It is shown in Figure 3.2.

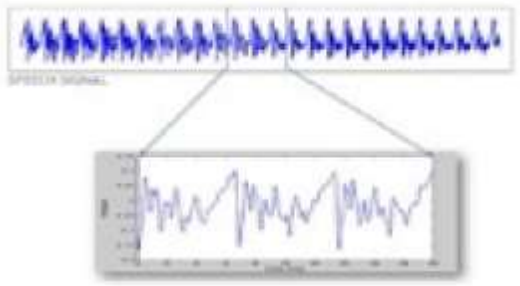

Figure 3.2 Framing the speech signal

Windowing - Windowing step is meant to window each individual frame, in order to minimize the signal discontinuities at the beginning and the end of each frame. Each frame is multiplied by an N sample window W (n). Here we use a hamming window. This hamming window is used to minimize adverse effects of chopping an N sample section out of the running speech signal. While creating the frames the chopping of N sample from the running signal may have a bad effect on the signal parameters. To minimize this effect windowing is done.

Figure 3.3(a) shows the widely used Hamming window and a single frame is multiplied by hamming window and the resulting signal is shown in Figure 3.3(b).

F

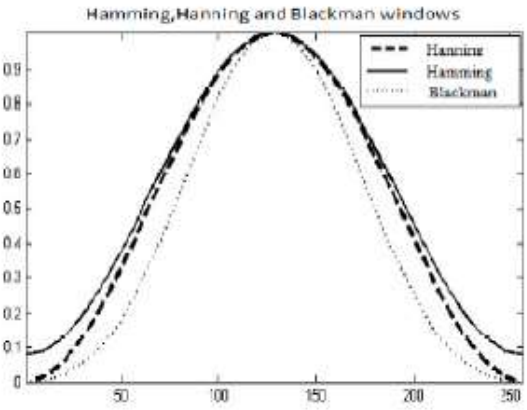

Fig 3.3(a) Hamming Window for speech signal

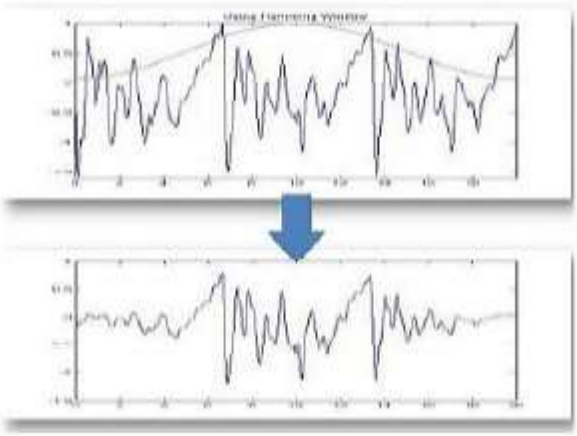

Fig 3.3(b) Windowing for speech signal

3.3 Feature Extraction

Features extraction means finding good parameters that helps to classify between the healthy and abnormal patients, features selection makes a boundary between each class.

3.3.1 MFCC Features

Mel frequency cepstral coefficients are given by:

These parameters are extracted by 32 filter bank applied on10 ms (256 points) Hamming windowed frames at 50% of overlap.

$$\tilde{C} = \sum_{k=1}^{K} \log(\tilde{S}) \cos[n(k - \tfrac{1}{2})] \frac{\Pi}{K} \quad (1)$$

3.3.2 Jittler and Shimmer

Jitter may occur during voice production, especially in vowel phonation, and it is defined as small fluctuations in glottal cycle lengths. Jitter and shimmer (amplitude perturbations) over successive speech cycles help give the vowel its naturalness in contrast to constant pitch and amplitude that can result in a machinelike sound. Moreover, jitter (and shimmer) contributes to the voice quality of a speaker. In terms of signal processing, jitter is a form of modulation noise. Specifically, jitter is a modulation of the periodicity of the voice signal. A high degree of jitter results in a voice with roughness that is usually perceived in recordings of pathological voices.

Therefore, a reliable estimation of jitter can be used to discriminate between healthy and dysphonic speaker.

Jittler: % change in cycle duration between cycles

Shimmer: % change in speech amplitude between cycles.

$$Jitter = \frac{\frac{1}{N-1}\sum_{k=1}^{N}\left|T_k - T_{k+1}\right|}{\frac{1}{N}\sum_{k=1}^{N}\left|T_k\right|} \quad (2)$$

$$Shimer = \frac{\frac{1}{N-1}\sum_{k=1}^{N}\left|A_k - A_{k+1}\right|}{\frac{1}{N}\sum_{k=1}^{N}\left|A_k\right|} \quad (3)$$

3.3.3 Formant Frequency

Formant is frequencies of resonance for each frame. It is often measured as an amplitude peak in frequency spectrum of the speech formants are resonances of the vocal tract. The formant frequencies are calculated using linear predictive coding and by finding the roots of prediction polynomial. The LPC finds the best IIR filter from the section of speech signal and then plots frequency response of filter.

3.3.4 Multi-Layer Spectral Estimation

Conventional spectral estimation of speech uses a Hamming window or a single taper. Using a single taper windowing result in a significant portion of the signal being discarded and the data points at the extremes being down-weighted, giving a high variance for the direct spectral estimate [12]. Hence, a multi-taper method is used so that the statistical information lost by using just one taper is partially recovered by using multiple windows for the same duration. The multi-taper spectrum is thus a weighted sum of the several tapered periodograms. Spectral estimation of a signal S using multi-taper method is as follows,

$$S_(m,k) = \frac{1}{M}\sum_{p=0}^{M-1}\lambda(p)\sum_{j=0}^{N-1} w_p(j)s(m,j)e^{-i2\pi\frac{k}{N}j} \quad (4)$$

where wp(j) is the pth data taper function. M is the number of tapers and lambda p is the weight corresponding to the pth taper, N is the

speech frame length and k are the FFT points. In practice, weights are designed so as to compensate for increased energy loss at higher order tapers.

3.3.5 Fundamental Frequency

The role of fundamental frequency F0 in the intelligibility of speech has been studied for both normal and dysarthric speech.[13] These studies suggest that a higher variation in F0 contributes significantly to increased intelligibility. However, for dysarthric speakers, the precision and flexibility of the vocal folds, articulators and other speech subsystems are lower, leading to reduced prosodic control, reflecting as a reduction in intelligibility. Additionally, studies show that the slower articulatory rate tends to be associated with low values of mean, maximum and variations of F0[14]. F0 measurements such as mean and variation are also indicative of the vocal loudness of speech, which has a bearing on speech intelligibility.

3.3.6 Noise to harmonic ration (NHR)

Noise-to-Harmonics ratio (NHR) is indicative of the abnormal vibratory characteristics of the vocal folds, manifesting as hoarseness in dysarthric speech. NHR is measured in dB, calculated by the ratio of noise energy or the aperiodic part of a sustained vowel to the energy of the periodic part. NHR can be used as a measure of voice quality and is defined as below

$$NHR(dB) = 10 * log\left(\frac{E_n}{E_p}\right) \quad (5)$$

where Ep is the energy of the periodic part and En is the energy of the noise. NHR has been used as one of discriminative features to evaluate the degree or severity of dysarthria [14]

3.3.7 Perceptual Linear Predictive and Log Perceptive Linear Predictive Coefficients

It is advancing version of Linear predictive coefficients (LPC) technique which emphasis the psychophysically based transformation. It remaps spectral features to bark scale as in contrast to Mel scale in MFCC to enhance middle hearing frequency range. Cube root approximation of loudness mimics the power law of hearing.

In LOG PLP spectral components are passed through band pass filter after taking logarithm on it in order to suppress additive distortion.

3.3.8 Harmonic to Noise Ratio (HNR)

this quantifies the ratio of signal information over noise due to turbulent air flow, resulting from an incomplete vocal fold closure in speech pathologies.

3.3.9 First and second derivatives of cepstral coefficient

these are useful to investigate the properties of the dynamic behaviour of the speech signal.

3.4 Classification Methods

3.4.1 Support Vector Machine

This is a discriminative classifier formally defined by a separating hyperplane that divides data belonging to different classes. The aim is to identify the class of belonging of the different data. Training a support vector machine requires the solution of a very large quadratic programming optimization problem. To resolve this problem the sequential minimum optimization (SMO) technique is used, which is able to divide the optimization problem into a series of smaller possible problems. The classification accuracy can be improved by selecting opportune form and parameters characteristic of the kernel function K (x, y).

The most popular kernel function forms are polynomial and radial basis one.

3.4.2 Decision Tree (DT): this technique is used to classify categorical data in which the learned function is represented by a decision tree. Decision trees are easy to interpret, capable of working with missing values and categorical and continuous data, characteristics of the medical field.

3.4.3Bayesian Classification (BC): this approach named after Thomas Bayes, who proposed the Bayes Theorem. The classify cation is achieved by evaluating the probabilistic model that represents a set of random variables and their conditional dependencies identified, respectively as nodes and strings. The major advantage is the easy interpretation of the results and the robustness in dealing with missing data.

3.4.4 Logistic Model Tree (LMT): this technique combines logistic regression models with tree induction. It consists of a standard decision tree structure with logistic regression functions at the leaves. Simple Logistic class implements this algorithm in WEKA.

3.4.5 Instance-based Learning algorithms: these algorithms use specific instances to achieve the classification predictions. The algorithms used are k-nearest neighbour where the classification is based on k nearest neighbours of a new instance and K* is an instance- based classifier that uses anentropy-based distance function to classify data.

3.4.6 Neural Network: The Generalization is the beauty of artificial neural network. It provides fantastic simulation of information processing analogues to human nervous system. Multilayer feed forward network with back propagation algorithm is the common choice in classification and pattern recognition. Hidden Markov Model, Gaussian Mixture Model, Vector Quantization are the some of the techniques for acoustic features to visual speech movement. Neural network is one of the good choices among all. Genetic Algorithm can be used with neural network for performance improvement by optimizing parameter combination.

Neural Network consists of input layer, hidden layer and output layer. Variable number of hidden layer neurons can be tested for best results. We can train network for different combinations of epochs with goal as minimum error rate.

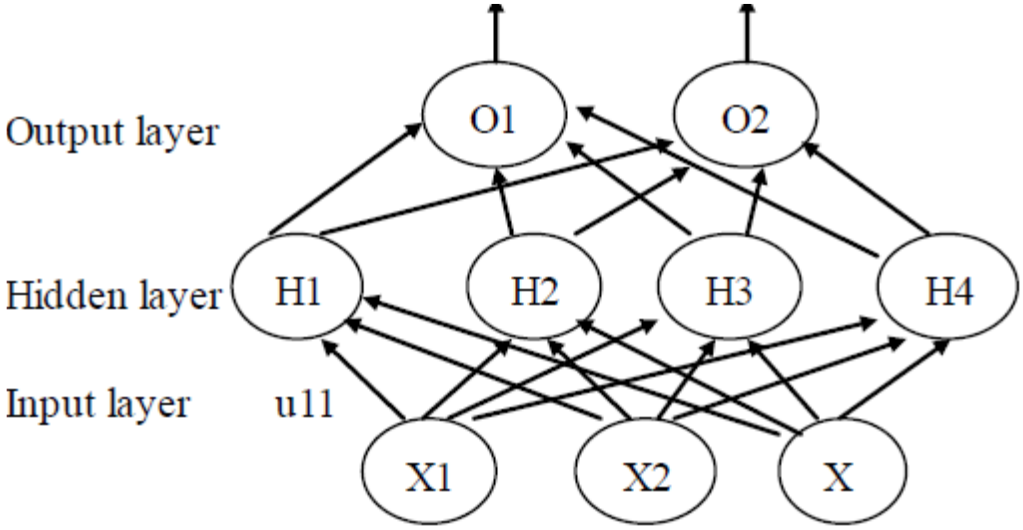

Fig-3.4 Structure of Neural Network

3.4.6 Multilayer Perceptron Neural Network (MLPNN): MLPNN is composed of three layers consisting of an input layer, one or more hidden layers and an output layer. The input layer distributes the inputs to subsequent layers. Input nodes have linear activation function and no thresholds. Each hidden unit node and each output node have thresholds associated with them in addition to the weights. The hidden unit nodes have nonlinear activation functions and the outputs have linear activation functions. The number of neurons in the hidden layer is dependent on the size of the input vector. The output layer has one neuron. This is used for classifying the pathological voice from the normal voice.

IV. Conclusion

A relative study is done to analyze the work done in the field of laryngeal pathologies identification. By analysis it shows that most of work done is based on acoustic features and classification methods. By using proper acoustic features and appropriate classification method, different laryngeal pathologies can be identified.

IV. Future Work

Future work should be included to use different machine learning techniques for classification of laryngeal pathologies and by using more feature combination more accuracy can be achieved.

Other than this, Disease severity can be found with the use of different machine learning algorithms.

V. References

[1] Mundt, J. et al "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," Journal of Neurolinguistics,20(1), 50–64, 2007.

[2] Alpert M. et al, "Reflections of depression in acoustic measures of the patient's speech", Journal of Affective Disorders,66, 59–69, 2001

[3] Ang J. et al, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog", ICSLP, 2002.

[4] Lee C.M., Narayanan S., "Towards detecting emotions in spoken dialogs", IEEE Transaction on Speech and Audio Processing, 2004.

[5] Chitralekha Bhat, Bhavik Vachhani, Sunil Kopparapu (2016). Recognition of Dysarthric Speech Using Voice Parameters for Speaker Adaptation and Multi-taper Spectral Estimation, INTERSPEECH, San Francisco, USA.

[6] LAURA VERDE, GIUSEPPE DE PIETRO AND GIOVANNA SANNINO (2018). Voice Disorder Identification by Using Machine Learning Techniques, IEEE.

[7] Lotfi Salhi, Talbi Mourad, and Adnene Cherif (2010). Voice Disorders Identification Using Multilayer Neural Network, The International Arab Journal of Information Technology, Vol. 7, No. 2.

[8] Eadie, Tanya, Doyle, Philip (2002). Classification of Dysphonic Voice: Acoustic and Auditory-Perceptual Measures, official journal of the Voice Foundation.

[9] V. Srinivasan, V. Ramalingam and P. Arulmozhi (2014). ARTIFICIAL NEURAL NETWORK BASED PATHOLOGICAL VOICE CLASSIFICATION USING MFCC FEATURES, International Journal of Science, Environment and Technology, Vol. 3, No 1, 2014, 291 – 302.

[10] João Paulo Teixeira, Paula Odete Fernandesa (2015). Acoustic Analysis of Vocal Dysphonia, Conference on enterprise Information Systems / International Conference on Project management / Conference on Health and Social Care Information Systems and Technologies.

[11] Mohamed FEZARI, Fethi AMARA and Ibrahim M. M. El-EMARY (2014). Acoustic Analysis for Detection of Voice Disorders Using Adaptive Features and Classifiers, Proceedings of the 2014 International Conference of Circuits, Systems and Control.

[12] G. A. Prieto, R. L. Parker, D. J. Thomson, F. L. Vernon, and R. L. Graham, "Reducing the bias of multitaper spectrum estimates, "Geophysical Journal International, vol. 171, no. 3, pp.1269–1281, 2007.

[13] R. Patel and P. Campellone, "Acoustic and perceptual cues to contrastive stress in dysarthria," Journal of Speech, Language, and Hearing Research, vol. 52, no. 1, pp. 206–222, 2009.

[14] K. Kadi, S. Selouani, B. Boudraa, and M. Boudraa, "Discriminative prosodic features to assess the dysarthria severity levels," in Proceedings of the World Congress on Engineering, vol. 3, 2013.