IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-8, Issue-1, Jan 2020 E-ISSN: 2347-2693

A Framework for Classification of Vocal Disorders without Clinical Intervention

Arpitha M.S.[1*], Nagarathna[2]

[1,2]Dept. of Computer Science, PES College of Engineering, Mandya, Karnataka, India

[*]*Corresponding Author: arpithabharadhwaj@gmail.com, Tel.: +91-7026578004*

DOI: https://doi.org/10.26438/ijcse/v8i1.7073 | Available online at: www.ijcseonline.org

Accepted: 12/Jan/2020, Published: 31/Jan/2020

Abstract— Voice disorders are abnormal characteristic of sound produced by larynx involving pitch, intensity, loudness. Nowadays Voice disorders are one among rapidly spreading diseases. Disordered quality of voice could also be a symptom for laryngeal diseases. The goal of this work is to build a model to identify the types of voice disorders that includes Normal, Dysphonia, Stammering and Vocal palsy. To deal with this classification problem, Machine learning classifier Support Vector Machine (SVM) is used. The results are evaluated in terms of accuracy, sensitivity, specificity and ROC based on the features extracted using Mel Frequency Cepstral Coefficients (MFCCs), they are the cepstral representation of audio clip.

Keywords— Voice disorders, Machine Learning, Classification, SVM, MFCC

I. INTRODUCTION

Voice disorders are severe medical condition mainly caused due to many risk factors includes allergies, abnormal tissue growth in laryngeal region, excess of alcoholic usage, improper clearing of throat, psychological stress etc., these leads to many problems such as nodules, polyps and sores on the vocal cords. It directly affects speech production [1]. In larynx there are two bands of muscle that vibrate to make sound. Voice problems are commonly found in children due to stress from excessive screaming or shouting. In case of actors or singers, loudness is generated without any damages to laryngeal system. Motor speech and swallowing disorders are the most common general vocal diseases. The sharpness of whisper or breathy whisper due to strained or contrasts with the weak are the main symptoms of vocal fold paralysis. To diagnose the pathological status of voice, Computer aided medical systems are being more used with lower cost. The most common symptom of conversion is "Aphonia (no voice)". In some patients the symptoms may reversible. In young children, "edema" is the most frequent vocal chords inflammation. Many of the traditional diagnostic methods depends on expensive devices and clinicians experience, which is cost effective and also causes delay for the patients who are in the place without the specialists and medical resources. Voice disorders assessment is carried out by speech language pathologist. The assessment process involves indirect laryngoscopy or videostroboscopy. Sometimes surgery is needed for voice disorders treatment. Examination for vocal disorder depends on patient's understanding, maturity or anatomy. Lumped element approach is the most common modelling framework in voiced speech investigations [2].

In past few decades a lot of research has been carried out related to automatic detection of voice pathologies. In most of the research work, Massachusetts Eye and Ear Infirmary are been used, the recordings of healthy and pathological voices are recorded in two different environments. Alternative techniques for the observation of vocal folds are direct or indirect laryngoscopy and video laryngoscopy [3]. These techniques are commonly used for the monitoring of larynx. Pattern recognition is also a special case in speech recognition. Speech is the most important communication channel which has prominent place plays in between human and machine. Glottic closure is the main parameter of vocal fold vibration. This vibratory characteristic is examined by stroboscopy. High speed endoscopy overcomes the limitations of stroboscopy to establish advancements in technology [4]. To discriminate between two different patient states namely diseased and non diseased, Receiver operating characteristic (ROC) analysis is used [5]. This technique is mainly used to quantify the accuracy for medical diagnostic tests. The basis for roc curve is separator curve, a pair of overlapping distribution forms the result for diseased and non diseased states. Accuracy indices is the most desirable property of ROC. It gives the trade off between true positive fraction and false positive fraction. Signal processing techniques are used to discriminate pathological voice disorders, this acts as a tool for video laringoscopy exams [6]. In socialization or in academic performance, children or any other individual with voice disorders experiences difficulties.

II. RELATED WORK

According to paper [7] vowels are recognised based on nonlinear speech parameters includes phase space anti-diagonal point distribution and maximal lyapunov exponent. In order to extract these features, reconstructed phase space is used. To obtain accurate description of speech linear filter is replaced with nonlinear models. Depending on the system, state of the phase can be finite or infinite or it can be a collection of all possible states. The result of nonlinear features shows discriminative power and combined vector features yield in increased accuracy.

The closure of glottis during the articulation of another voice or production of creaky sound while speaking some other speech is highlighted [8]. By introducing different prephonatory configurations, normal voice can be altered. Vocal hyper function(VH) is a chronic condition due to abnormal functioning of muscles. Vocal hyper function may lead to chronic tissue trauma, it is also known as phonotrauma. This leads to deposition of more trauma tissues and also formation of lesion. In this paper, a triangular glottal shaped model is proposed for vocal folds. To propagate the acoustic pressure of glottis, an algorithm called, wave reflection analog is proposed.

Laryngeal disorders are another name for the disordered quality of voice. Auditory-perceptual evaluation is the primary process to detect the quality of voice [9]. To calculate power spectrum in MFCC feature extraction, Multitaper Spectrum estimation process is carried out. The difference in estimated and actual spectrum reduces the bias spectrum. Gaussian Mixture Model (GMM) approach is used to model this system.

Vocal cord paralysis is one the serious voice disorders which leads to the paralysis to vocal region, this condition is also known as Reinke's edema. To diagnose this severe medical condition, a novel method is produced using convolutional deep belief networks [10]. This system uses recordings of normal and pathological speech samples as input to the network. In order to pre-train the convolutional neural network, convolutional deep belief networks is used. For the analysis of performance, real voice samples from Saarbrucken Voice database are considered.

The coupling between the aerodynamic forces and the tissue parameters produces vocal fold vibrations. The vocal chord can be paralysed or partly paralysed. This is the reason to generate spectrum of acoustical sound that influences both fluid flow and also vocal fold dynamics. Since voiced speech involves more coupled interactions, it is a complex process. The regulation of lung pressure generates sound which is known as voice box where, the larynx is present in the anterior portion of neck. Vocal fold structure is modelled by a collection of discrete coupled mass spring damper system and acoustic loading function. Vocal fold is housed by larynx and it is the most less active ventricular fold.

III. PROPOSED METHOD

The idea is to develop a machine learning system where the voice signals can be acquired by a microphone or smart phone or tablet, which is processed in real time to extract the voice features and it is analyzed by using machine learning classifier to detect the presence or absence of voice disorders, as shown in Figure 1.

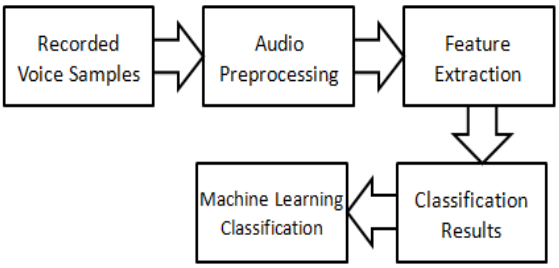

Figure 1: The flow chart of classification of vocal disorders

Recorded Voice Samples are the input for this work. It consists of four different types of voice categories such as Dysphonia, Stammering, Normal and Vocal Palsy. The length of each signal is 10 seconds, which are recorded at comfortable level of loudness i.e. the distance between mouth and microphone is kept at the distance of 15-20 cm and is saved in uncompressed wave form audio file format(.wav).

Audio Preprocessing is the preliminary process of this work where, audio signals are sound waves which travel through air. It can either be represented both in digital format and in analog format where, analog format operate on electrical signal and digital format operate on mathematical notations. The audio waveform in its digital form is represented in terms of binary numbers. In this, the prerequisite step is to remove the noise in the recorded voice samples since sound consists of audible variation in air. In order to carry out audio preprocessing, there are several steps to be followed. They are as follows,

1. Load audio files
2. Extract sample rate and Nyquist frequency from audio samples
3. Fix pass band and stop band frequencies for the audio samples
4. Apply median filter
5. Plot magnitude spectrum
6. Obtain smoothed signal

The general sampling rate of a voice signal is 44,100 Hz with a resolution of 16-bit and a sequence of 44,100 points in time

series represent one second of a person's voice. Nyquist frequency is the minimum rate at which a signal can be sampled and it is taken as half of the sample rate. Pass band is the range of frequencies that can pass through a filter and stop band does not allow the pass band frequency. Median filter is a non-linear filter used for the reduction of noise to convert the signal into second order section for stability. Plotting of magnitude spectrum is necessary to notify the noisy regions which gives normalized frequency(π rads/sample). In smoothing, the data points of a signal are modified so that individual points higher than the adjacent points are reduced and points that are lower than the adjacent points are increased leading to a smoothed signal.

Feature Extraction is the process of extracting the necessary features from the input voice samples. The speech feature extraction is about reducing the dimensionality of input vector. For this process, Mel Frequency Capstral Coefficients(MFCC's) are computed to obtain the required outcome :

1. Frame the signal into short frames of length 20 milliseconds
2. For each frame, calculate the estimate of power spectrum
3. Apply mel filterbank to the power spectra and sum the energy in each filter
4. Take the logarithm of all filtered energies
5. Take the discrete cosine transform of the log filterbank energies and obtain the MFCC's

To divide the speech signal into frames, Hamming window function (HMW) is used at fixed intervals of 20ms that are statically stationary. HMW is sometimes called as raised cosine window. This is used because it provides edge effect, which removes silence part from both starting and ending region of all the frames. Windowing is done mainly because to avoid truncation of the signal. Power spectrum describes the distribution of power values as a function of frequency, where power is the average of signal. It is computed for the entire signal using periodogram function. Mel filterbank is applied for the calculated spectral power of each frame which consists of series of overlapping triangular filters defined by their center frequencies $lf_{c(m)}$. The parameters that define Mel filterbank are number of Mel filters(26) minimum frequency lf_{min} and maximum frequency lf_{max}. Usually MFCC,s are calculated in single window. The coefficients are derived from the Fourier Transform of the audio clip. The features extracted for calculation are first quartile, third quartile, median, minimum and maximum of each row with n data point of the resulted matrix. For example, if a cleaned voice sample is five-second long, the time series has 5*44,100 = 220,500 points, and there are 5/0.020 = 250 frames if 0.020 second is the frame length. The resulted dataset is a matrix of shape 26 X n, where 26 being the number of filters and n being the total number of voice samples.

Classification is to classify the input voice samples, machine learning classifier Support Vector Machine(SVM) is used. SVM defines the hyperplane to classify binary classes. It searches for optimal hyperplane or decision boundaries. Support vector machine is based on supervised learning method. The training samples in support vector machine is separable by hyperplane and it is computed by decision function f(x) = sign(w . x) + b where, w is a weighted vector and b is a threshold cut-off. The optimal hyperplane can be defined in infinite number of different ways. Hence, for this work Multi SVM is chosen, which defines more than two hyperplanes to classify the disordered voice samples into different classes. The kernels used in SVM are linear kernel, Gamma kernel and polynomial kernel. Before applying Support vector machine algorithm, the data should be standardized. It consists of two datasets X-> samples and Y->classes, line space is created to define the hyperplane and plotting of hyperlines to distinguish between the different classes. After applying the multi support vector machine algorithm, the classes of disorders that need to be classified are dysphonia, stammering, normal and vocal palsy based on the defined value of the hyperplanes.

IV. CONCLUSION

In the present work, voice disorder detection system is to be developed, which deals with the prediction and classification of four different types of vocal disorders using machine learning classifier SMV. The parameters of feature extraction can be used to detect other set of vocal disorders apart from the proposed work. In order to improve the accuracy of the system, other classifiers can also be applied.

REFERENCES

[1] N. Souissi and A. Cherif, "Dimensionality reduction for voice disorders identification system based on mel frequency cepstral coefficients", In the Proceedings of the 2015 7th International Conference on Modelling, Identification and Control, pp. 1-6, 2015.

[2] Byron D. Erath, Matias Zanartu, Kelly C. Stewart, Michael W. Plesniak, David E. Sommer, Sean D. Peterson, "A review of lumped-element models of voiced speech", Speech Communication Publisher, US, pp. 667-690, 2013.

[3] Arias-Londoño JD, Godino-Llorente JI, Markaki M, Stylianou Y, "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices", Logoped Phoniatr Vocol, 2011 Jul, pp. 36(2):60-9.

[4] Edwin L, Kendig, Robert W. Wilmott, Victor Chernick, "Disorders of the Respiratory Tract in Children", Elsevier Health Sciences Publisher, 2012.

[5] K. Hajian-Tilaki, "Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation", Caspian journal of internal medicine, vol. 4, no. 2, p. 627, 2013.

[6] S. C. Costa, B. G. Aguiar Neto and J. M. Fechine, "Pathological voice discrimination using cepstral analysis, vector quantization and

Hidden Markov Models", 2008 8th IEEE International Conference on BioInformatics and BioEngineering, Athens, 2008, pp. 1-5.
[7] Fathima Kunhi Mohamed, Lajish V.L, et al. "Nonlinear Speech Analysis and Modeling for Vowel Recognition", 6th International Conference On Advances In Computing & Communications, 2016, Cochin, India.
[8] Galindo, Gabriel E, et al, "Modeling the pathophysiology of phonotraumatic vocal hyperfunction with a triangular glottal model of the vocal folds". Journal of Speech, Language, and Hearing Research, Vol. 60 2452–2471, 2017
[9] J. I. Godino-Llorente, P. Gomez-Vilda and M. Blanco-Velasco, "Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters", in IEEE Transactions on Biomedical Engineering, vol. 53, no. 10, pp. 1943-1953, 2006.
[10] Wu, Huiyi, et al. "A deep learning method for pathological voice detection using convolutional deep belief networks", Interspeech 2018.
[11] K. Umapathy, S. Krishnan, V. Parsa and D. G. Jamieson, "Discrimination of pathological voices using a time-frequency approach", in IEEE Transactions on Biomedical Engineering, vol. 52, no. 3, pp. 421-430, March 2005.
[12] T. Marciniak, R. Weychan, S. Drgas, A. Dabrowski, and A. Krzykowska, "Speaker recognition based on short polish sequences", in Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings (SPA), 2010. IEEE, 2010, pp. 95–98.
[13] M. S. Hossain, "Patient state recognition system for healthcare using speech and facial expressions", Journal of medical systems, vol. 40, no. 12, p. 272, 2016.
[14] International Journal of Scientific Research in Network Security and Communication (ISSN: 2321-3256).
[15] Gourish Malage, Kiran Kumari Patil, "A Voice based Farmer Information System", International Journal of Scientific Research in Computer Sciences and Engineering, Vol.7, Issue.6, pp.220-224,2019.
[16] Sujitha Perumal, Mohammed Saqib Javid, "Voice Enabled Smart Home Assistant for Elderly", International Journal of Scientific Research in Computer Sciences and Engineering, Vol.7, Issue.11, pp. 30-37, 2019.
[17] M. Mat Baki, G. Wood, M. Alston, P. Ratcliffe, G. Sandhu, J. Rubin, and M. Birchall, "Reliability of operavox against multidimensional voice program (mdvp)", Clinical Otolaryngology, vol. 40, no. 1, pp. 22–28, 2015.

Authors Profile

Miss. Arpitha M.S. Pursuing Master of Technology in Computer Science and Engineering from PES College of Engineering, Mandya. She has received her Bachelor's degree in Computer Science and Engineering from Academy for Technical and Management Excellence (ATME), Mysuru.

Dr. Nagarathna currently working as Professor in Department of Computer Science and Engineering, PES College of Engineering, Mandya.

