2018 Second International Conference on Advances in Electronics, Computer and Communications (ICAECC-2018)

Classification of Healthy and Pathological voices using MFCC and ANN

Smitha
Department of CSE
NMAM Institute of Technology, Nitte
smitha.k.rai@gmail.com

Sarika Hegde
Department of CSE
NMAM Institute of Technology, Nitte
sarika.hegde@yahoo.in

Surendra Shetty
Department of MCA
NMAM Institute of Technology, Nitte
hsshetty4u@yahoo.com

Thejaswi Dodderi
Audiologist & Speech Language Pathologist
Nitte Institute of Speech & Hearing, Mangaluru
thejaswi@nitte.edu.in

Abstract— **The automatic system for classification of healthy and pathological voices has received a significant attention in the research of early detection and diagnosis of voice disorders. In this work, we propose a method to classify the healthy and pathological voices. To implement this system, we use audio recordings of normal and pathological voices. We extract Mel Frequency Cepstral Coefficients (MFCC) from the voice signals and use a visualization technique to explore the capability of these features in discriminating healthy and pathological voices. In this study, we use Artificial Neural Network (ANN) to classify the extracted features. Here, we present the results of experiments with varying number of neurons in the hidden layer and also with various frame sizes. The best obtained accuracy is 99.96%.**

***Index Terms*—Voice Disorder, Artificial Neural Network, classification, Pathological voices.**

I. Introduction

Voice is the natural medium of communication among human beings. Vocal communication is an essential skill required by human beings for expressing their feelings, earning livelihood and in other day to day social interactions. Some professionals like singers, actors, politicians, and teachers mainly depend on their voice for their livelihood. Nowadays voice disorders are increasing significantly, due to hearing loss, neural disorders, brain injury, drug exploitation, physical injuries, vocal misuse, and unhealthy social habits, etc. Vocal fold disorders can also be caused as a result of extreme usage of the voice. A person with a voice disorders may experience uneasiness, awkwardness, depression, and will face many problems while communicating with other people. It may lead to many social and personal complications. So, early detection of voice disorder through automatic voice disorder detection system plays an important role. By detecting the voice disorder in early stage, one can visit speech pathologist and can cure their voice complications.

The main objective of this proposed method is to classify the healthy and pathological voices by way of developing an automatic voice disorder detection system.The main purpose of this system is to help the patients in identifying the voice disorders early so as to consult Ear Noise Tongue (ENT) or voice therapist. The remaining part of this paper is organized as follows: Section 2 gives you some of the related works, section 3 provides the methodology, section 4 presents the experiments and results and finally, section 5 draws a conclusion.

II. Literature Survey

In recent years, many automatic voice disorder detection systems have been developed using various databases, different feature extraction techniques, and various classification algorithms.

The features like, F0, F1, F2, F3 formants frequency levels, Harmonic to Noise Ratio (HNR), jitter, shimmer, intensity was analyzed in paper [1]. Vocal fold disorder detection system have been developed using MFCC and Gaussian Mixture Model (GMM) in [2]. They have tested the system using 12, 24 and 36 MFCC with 4,8,16 and 32 GMM and obtained 91.66% accuracy. In [3], the authors have proposed a new feature which is based on wavelet packet decomposition and MFCC. Here ANN is used as a classifier and achieved 91.54% accuracy.

In [4], the authors have investigated the methods of acoustic voice analysis. Here, MFCC is used and are modeled by GMM. They have got the best accuracy with 39 coefficients. In work [5], MFFC have been used along with feature space transformation technique. The influence of spectral envelope features, while calculated with a high order LPC has been studied in [6-7] and found that the frequency and bandwidth of the first peak obtained from LPC of 30th order were a valuable feature for voice pathology detection.

In [8], voice pathology detection system using Modified Mellin Transform feature and ANN classifier with 10 hidden layers have been proposed and achieved 96.48% accuracy. In study [9], proposed a voice disorder analysis system using glottal signal parameters with Multilayer Perceptron and obtained 96.6% accuracy. Some works have used variants of ANN to classify normal and pathological voice [10-11]. The

978-1-5386-3785-2/18/$31.00 ©2018 IEEE

Support Vector machine (SVM) also have been used as a classifier in many research works efficiently [12-16].

III. METHODOLOGY

The main purpose of this work is to classify the healthy and pathological voices. MFCC is used for feature extraction because it is found that it extracts the useful information from the human voice which will accurately classify the healthy and pathological voices. The feature vector is created with 19 coefficients. This feature vector is fed into a Neural Network. The Scaled Conjugate Gradient (SCG) back propagation method is used to train the network. Then this method is evaluated using Mean Squared Error (MSE) and Percent Error (%E) The methodology of the proposed system is as shown in Fig. 1.

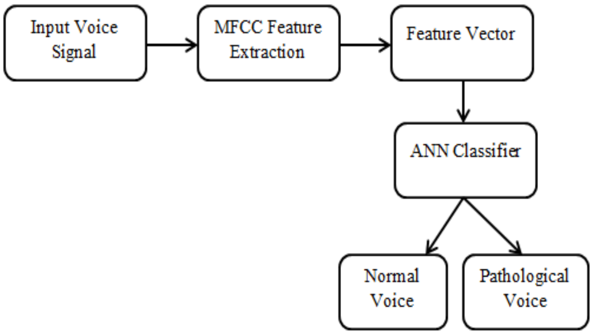

Fig. 1. Methodology of Proposed system

A. Data Description

The database was provided by Speech Language Pathologist from the Nitte Institute of Speech and Hearing Mangaluru. The dataset contains both healthy and pathological voices with the recordings of sustained vowel /a/. All the voices are recorded at 16kHz frequency and are in .wav format. Each file is divided into frames and each frame is considered as one sample.

The spectrum of healthy and pathological voice signals are shown in Fig. 2 and Fig. 3 respectively.

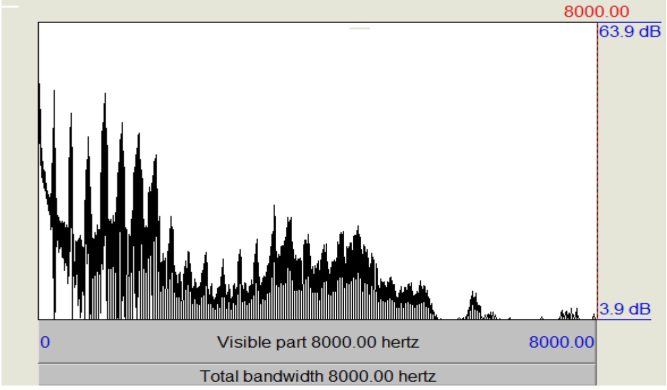

Fig. 2. Spectrum of a normal voice

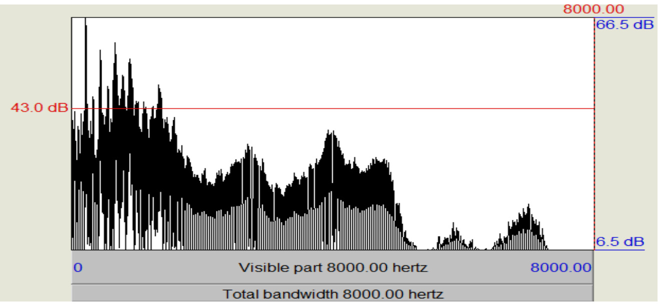

Fig.3. Spectrum of a pathological voice

A. Mel Frequency Cepstral Coefficients (MFCC)

The Mel Frequency Cepstral Coefficients (MFCC) are very popular features used for representing speech signals. This feature seeks to emulate human perception of decoding the speech signal to extract the information from the input. The steps involved in extracting the MFCC features [17] are shown in Fig. 4.

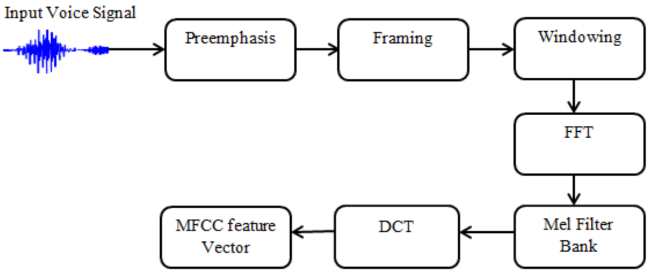

Fig. 4. MFCC feature Extraction

To extract the MFCC feature, the speech signal is divided into frames by applying framing, windowing and overlapping techniques. Then for each frame, Fast Fourier Transform (FFT) algorithm is applied to calculate the Discrete Fourier Transform (DFT) coefficients. Then the resulting power spectrum is converted to Mel Frequency scale. This is done by creating a Mel filter bank. An approximate conversion between a frequency value in Hz (*f*) and mel is given by (1) [18].

$$\text{mel(f)} = 2595\log_{10}\left(1+\frac{f}{700}\right) \qquad (1)$$

Mel scale is used because it is logarithmic and emulate the way in which filtering is done in human ears. The resulted power spectrum is filtered using this triangular filterbank constructed with Mel-scale. Then the coefficients can be obtained from the filtered spectrum by taking the logarithm of sub band energies followed by the Discrete Cosine Transform (DCT). The equation for computing DCT coefficients is given by (2).

$$C_i = \sum_{k=1}^{K} (\log(S_k)).\cos\left(\frac{i\Pi}{\text{K}}\left(k-\frac{1}{2}\right)\right); i = 1,2,\ldots,K \qquad (2)$$

Where C_i is the i^{th} MFCC, S_k is the output of k^{th} filterbank and K is the number of filter banks. This final representation is an approximation of the compressed and equalized signal produced by the mechanism of human hearing. This feature makes an attractive feature for context-aware research, and has been found to be very discriminatory [19].

Using this procedure, we have extracted 19 MFCC coefficients which depend on the number of filters used in filter bank. All the 19 coefficients have been used in this work.

B. *Artificial Neural Network (ANN)*

From the literature, it has been found that, an ANN is one of the efficient techniques used for classification of healthy and pathological voices [3, 8-11, 20-21]. So, in this paper, we have used neural network with one hidden layer to classify the extracted features. An ANN is made up of the number of interconnected neurons. The neurons are the processing elements. An ANN contains at least one input layer and one output layer. It can also have hidden layer between the input and output layer. This is called as Multilayer Neural Network (MNN). The number of hidden layers can be varied to achieve the classification efficiency. Each neuron in the one layer is interconnected with all the neurons of the next layer. The organization of a neuron is shown in Fig. 5.

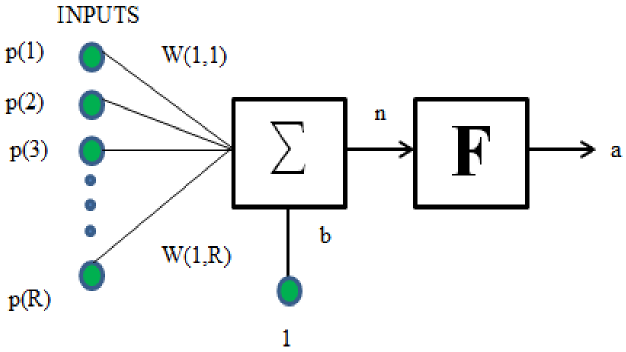

Fig. 5. Organization of a Neuron

A neuron consists of links, linear combiner and an activation function. Each link contains a weight factor ‘w’. Each input is multiplied by a weight factor and all the weighted inputs are added together by linear combiner. The bias ‘b’ is also added to the weighted sum. The output mainly depends on the activation function. The sigmoid activation function is used in this work. In multilayer network, the output of the one layer will be fed into the next layer as input. This continues until the output layer is reached.

Once the output is obtained, the output will be subtracted with the target to calculate the error. The error rate can be reduced through back propagation method. This process stops, when the rate of change of error becomes minimal. SCG back propagation method is used in this work.

IV. Experiments and Results

The experiments are simulated in the MATLAB 2015a software. In this work, the voice signals are converted into frames using framing, windowing, and overlapping techniques. In our experiments, each frame is considered as one element. MFCCs are extracted from each frame as per the MFCC feature extraction procedure. To visualize the feature vector, we have randomly selected the 100 elements from both Normal and Pathological voices and plotted first 3 coefficients from each feature vector as shown in Fig. 6.

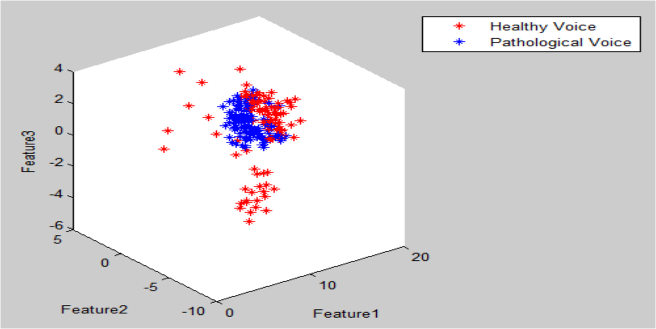

Fig. 6. Visualization of MFCC Feature

After creating the feature vector, a neural network is created built as shown in below Fig. 7. The feature vector (input data) and target data are fed into the network.

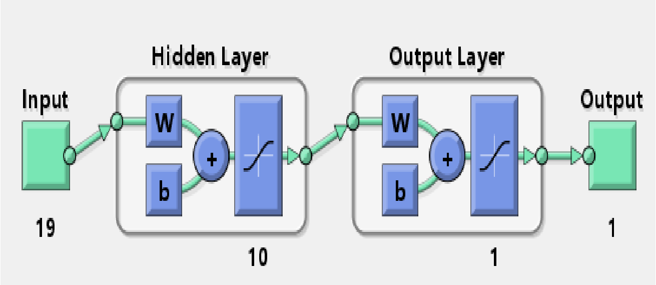

Fig. 7. A Neural Network

We have divided our dataset into 75% training data, 15% validation data, and 15% test data. The network is trained using SCG back propagation method. The neural network is evaluated using Mean Squared Error (MSE) and Percent Error (%E). Since the data is divided randomly, we have trained the network 20 times and have taken the average MSE and %E.

Initially, to investigate the number of neurons in the hidden layer required to get the optimal result, the neural network is built with a varying number of neurons (from 1 to 10 neurons). The frame size is taken as 256 and overlapping as 100 for these experiments. The obtained MSE and %E for varying number of neurons is presented in Table. I.

In this experiment, it has been observed that, the network with 9 neurons in the hidden layer got the less %E compared to other networks. It also have been observed that, %E is less than 1% for all cases. The number of neurons versus %E graph is as shown in below Fig. 8.

TABLE I. COMPARISON OF MSE AND %E FOR VARYING NUMBER OF NEURONS

No. of Neurons	MSE	%E
1	2.83e-03	2.95e-01
2	2.19e-03	2.18e-01
3	5.05e+01	3.33e-01
4	2.16e-03	2.31e-01
5	1.52e-03	1.79e-01
6	1.46e-03	1.79e-01
7	1.37e-03	1.35e-01
8	1.26e-03	1.47e-01
9	**1.05e-03**	**1.28e-01**
10	1.64e-03	2.12e-01

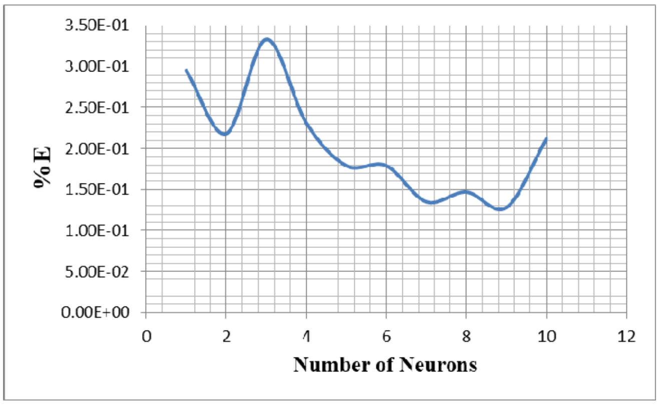

Fig. 8. Number of neurons vs. %E.

We have also experimented the network with various frame sizes and overlapping. Initially we have taken the frame size as 256 and overlapping as 100. Then the frame size is changed to 512, 1024, 2048, 4096 and the overlapping is also changed accordingly. The number of neurons in the hidden layer is taken as 9 for all the cases. The obtained MSE and %E for various frame sizes are given in below Table. II.

TABLE II. COMPARISON OF ACCURACY WITH DIFFERENT FRAME SIZE

Frame Size	MSE	%E	Accuracy (%)
256	1.05e-03	1.28e-01	99.87
512	**4.34e-04**	**3.86e-02**	**99.96**
1024	7.80e-04	7.73e-02	99.92
2048	2.25e-03	2.06e-01	99.79
4096	6.51e-03	7.29e-01	99.27

The accuracy of network with varying frame sizes (varying number of observations) is plotted as shown in the following Fig. 9. Here, it has been observed that the MSE and %E are varied with different frame sizes. From the result, it has been noted here that as the frame size increases, the accuracy decreases. The least %E is for frame size 512 and accuracy is 99.96%.

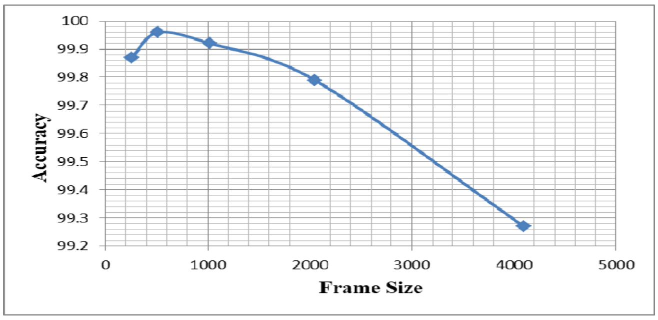

Fig. 9. Frame size vs. Accuracy

V. CONCLUSION

The method for classification of healthy and pathological voices is proposed in this paper. MFCC is used for creating the feature vector. The visualization of MFCC feature is also presented. The feature vector with 19 MFCCs are fed into ANN for classification. The dataset is divided into 75% training data, 15% validation data, and 15% test data. First, the number of neurons in the hidden layer is investigated and found that least %E (i.e. 1.05e-03) is obtained by a neural network with 9 neurons in the hidden layer. In the second experiment, the network is tested for different frame sizes and it has been observed that the best accuracy 99.96% (i.e. 3.86e-02%E) is obtained by the frame size 512.

VI. ACKNOWLEDGEMENT

This work is supported by the Nitte Institute of Speech and Hearing, Mangaluru. We are thankful to the Speech Language Pathologist from this Institute for providing the database used in this work.

REFERENCES

[1] Sonu, R. K. (2012). Sharma,"Disease detection using analysis of voice parameters,". Int. J. Comput. Sci. Commun. Technol, 4(2).

[2] Ali, Z., Alsulaiman, M., Muhammad, G., Elamvazuthi, I., & Mesallam, T. A. (2013, November). Vocal fold disorder detection based on continuous speech by using MFCC and GMM. In GCC Conference and Exhibition (GCC), 2013 7th IEEE (pp. 292-297). IEEE.

[3] Majidnezhad, V., & Kheidorov, I. (2013). An ANN-based method for detecting vocal fold pathology. arXiv preprint arXiv:1302.1772.

[4] Fezari, M., Amara, F., & El-Emary, I. M. (2014, February). Acoustic Analysis for Detection of Voice Disorders Using Adaptive Features and Classifiers. In Proc. 2014th Int. Conf. on Circuits, Systems and Control, Switzerland, Feb. 22–24 2014.

[5] Arias-Londoño, J. D., Godino-Llorente, J. I., Sáenz-Lechón, N., Osma-Ruiz, V., & Castellanos-Domínguez, G. (2010). An improved method for voice pathology detection by means of a HMM-based feature space transformation. Pattern recognition, 43(9), 3100-3112.

[6] Cordeiro, H. T., Fonseca, J. M., & Ribeiro, C. M. (2013). LPC spectrum first peak analysis for voice pathology detection. Procedia Technology, 9, 1104-1111.

[7] Cordeiro, H., Fonseca, J., & Meneses, C. (2014, August). Spectral envelope and periodic component in classification trees for pathologial voice diagnostic. In Proceedings of 36thAnnual IEEE International Conference of Engineering in Medicine and Biology Society (EMBC), (pp. 4607-4610).

[8] Francis, C. R., Nair, V. V., & Radhika, S. (2016, October). A scale invariant technique for detection of voice disorders using Modified Mellin Transform. In proceedings of International Conference on Emerging Technological Trends (ICETT), (pp. 1-6).

[9] Kohler, M., Vellasco, M. M., & Cataldo, E. (2016). Analysis and classification of voice pathologies using glottal signal parameters. Journal of Voice, 30(5), 549-556.

[10] Hariharan, M., Polat, K., & Sindhu, R. (2014). A new hybrid intelligent system for accurate detection of Parkinson's disease. Computer methods and programs in biomedicine, 113(3), 904-913.

[11] Srinivasan, V., Ramalingam, V., & Arulmozhi, P. (2014). Artificial Neural Network Based Pathological Voice Classification Using MFCC Features. Int. J. Science, Environment Technology, 3(1), 291-302.

[12] Saidi, P., & Almasganj, F. (2015). Voice disorder signal classification using m-band wavelets and support vector machine. Circuits, Systems, and Signal Processing, 34(8), 2727-2738.

[13] Al-nasheri, A., Ali, Z., Muhammad, G., Alsulaiman, M., Almalki, K. H., Mesallam, T. A., & Farahat, M. (2015, February). Voice pathology detection with MDVP parameters using Arabic voice pathology database. In proceedings of 5th National Symposium on Information Technology: Towards New Smart World (NSITNSW), (pp. 1-5).

[14] Muhammad, G., Ali, Z., Alsulaiman, M., & Al-Mutib, K. (2014). Vocal fold disorder detection by applying LBP operator on dysphonic speech signal. Proc. Recent Adv. Intell. Control Model. Simul, 222-228.

[15] Muhammad, G., Altuwaijri, G., Alsulaiman, M., Ali, Z., Mesallam, T. A., Farahat, M., & Al-nasheri, A. (2016). Automatic voice pathology detection and classification using vocal tract area irregularity. Biocybernetics and Biomedical Engineering, 36(2), 309-317.

[16] Muhammad, G., Alsulaiman, M., Ali, Z., Malki, K., Mesallam, T., & Farahat, M. (2017). Voice Pathology Detection and Classification using Auto-correlation and entropy features in Different Frequency Regions. IEEE Access.

[17] Slaney, M. (1998). Auditory Toolbox: a Matlab toolbox for auditory modeling. Work Technical Report, Interval Research Corporation, 29-32.

[18] Peltonen, V., Tuomi, J., Klapuri, A., Huopaniemi, J., & Sorsa, T. (2002, May). Computational auditory scene recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1941-1944, Orlando, Fla, USA, May 2002.

[19] Rabiner, L. R., & Juang, B. H. (2006). Speech recognition: Statistical methods. Encyclopedia of Language & Linguistics (2nd Ed), 1-18.

[20] Linder, R., Albers, A. E., Hess, M., Pöppl, S. J., & Schönweiler, R. (2008). Artificial neural network-based classification to screen for dysphonia using psychoacoustic scaling of acoustic voice features. Journal of voice, 22(2), 155-163.

[21] Godino-Llorente, J. I., & Gomez-Vilda, P. (2004). Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. IEEE Transactions on Biomedical Engineering, 51(2),380-384.