Voice Pathology Identification using Deep Neural Networks

C.S.Kanimozhiselvi, M.Balaji Prasath, T.Sathiyawathi

Abstract— ***The human voice construction is a complex biological mechanism capable of Changing pitch and volume. Some Internal or External factors frequently damage the vocal cords and change quality of voice or do some alteration in the voice modulation. The effects are reflected in expression of speech and understanding of information said by the person. So it is important to examine problem at early stages of voice change and overcome from this problem. ML play a major role in identifying whether voice is pathological or normal in nature. Voice features are extracted by Implementing Mel-frequency Cepstral Coefficients (MFCC) method, and examined on the Convolutional Neural Network (CNN) to identify the category of voice.***

Keywords : Classification, Convolutional Neural Network, MFCC, Voice disorder.

I. INTRODUCTION

When the damage occurs at the vocal cord, it affects the production of voice quality and hence the generated voice called as pathological voice. Voice pathologies affect the larynx and result in uneven vibrations of the vocal folds. Poor voice quality can affect the individual's ability to communicate both socially as well as in the place of work, thus affects the quality of life, and it has a major impact on economy considering the costs of medical diagnosis and treatment. Voice disorders can be classified as: organic, functional.[1] along with several methods to examine voice disorders, the acoustic analysis have confirmed its effectiveness. Normal voice is produced at larynx and generates good quality of speech sounds.[2] Sometimes, many violent speech, normally called as vocal hyber function, produce a voice disorders .[3] The objective of this paper is to automatic identification of one of the communicative disorder called voice disease based on voice database given to the Deep Neural network. Here we have used our own database of audio files collected from the schools (Erode District, Tamil Nadu) as an input to the neural network model.

LITERATURE REVIEW

Thongluan Laosaphan et al [4], [1] proposed a voice identification by using coefficients extracted from voice signal of spoken words based on the principle of MFCC feature extraction for performing utterance recognition. This method provides better recognition rates when training the words with SVM and provides the improved performance.

Gaganpreet Kaur et al [6] examined the research done in the domain of speaker recognition. The different methods used for feature extraction and feature classification had been discussed. Some methods preferred over others such as MFCC for feature extraction had better performance rather than LPC, because MFCC were most consistent with human hearing due to Mel scale representation. Thus, it was concluded that feature extraction of GMM performs better as they require fewer amounts of data to train the classifier. It also decreases the memory usage of the system.

Aman Ankit et al [5] introduced ASR techniques and had put forth some of the essential information. The voice Recognition System and other methods implemented in ASR developed for various languages. HMM and HMM Toolkit had been used. It describes the methods used and comparative study of the performance system developed.

Daria pane et al [7] proposed a multi dimensional system that can classify voices between pathological and normal. In this work, 28 voice features is examined using PCA, KPCA and NLPCA in some pathological detection methods implemented by testing the A, E, I, O and U vowels, at a high, low and normal pitch. The outcomes from different methods were implemented at 10-fold cross-validation. The outcomes show that, the PCA Techniques used to improve the data and we still have a 90% of the variance.

Al-nasheri, A Zulfiqar, [6] implemented a Robot Arm that is used to pick and place an object via recognition of speech sounds. This method implemented at python 2.7 version and works well for all speech sounds and recognize better. The outcomes obtained from the speech recognition system has a high average accuracy rate of speech recognition, which is 80% of the respondents trained data and 70% of the respondents for not trained data.

Vimala.C; Dr.V.Radha et al [7] presented a voice pathology identification system to identify the patient voice quality. The Speech signals are sampled at several samples, and CNN models are implemented to process those speech samples at several layers of neurons.

Manuscript published on November 30, 2019.

* Correspondence Author

Dr.C.S.Kanimozhiselvi*, Computer Science and Engineering, Kongu Engineering College, Perunduari, Erode, Tamil Nadu, India. (Email: kec.kanimozhi@gmail.com)

Mr.M.Balaji Prasath, Computer Science and Engineering, Kongu Engineering College, Perunduari, Erode, Tamil Nadu, India. (Email: itprasath@gmail.com)

Miss.T.Sathiyawathi, Computer Science and Engineering, Kongu Engineering College, Perunduari, Erode, Tamil Nadu, India.

Then, a fusion approach implemented to fuse the features from the CNN models. Another direction is the use of different types of inputs, such as voice and EGG signals, combined by deep fusion strategy.

Björn W.Schullere et al [12] proposed a deep-learning based methods to differentiate between unconditional speeches from steady speech and examines its efficiency and utility compared with other data mining algorithms.The results evaluated on Several Speech database shows that, DNN perform better than traditional GMM and SVM in steady increase in detection accuracy based on three representative features.

Hou, J.-C.; Wang, S.-S.; Lai [11] presents the deep learning model that can classify the speech sounds based on the generated spectrograms.Ecological sounds of spectrogram images used to train the CNN model and tensor deep stacking network that are proposed to used in sound classification application.

III. AUDIO DATA PROCESSING

These sound files are converted into digital wave form represented as .wav format. These files are sampled at the rate of 44.1khz. Samples are done at discrete intervals called as sampling rate. Each sample is the amplitude of the wave at a particular time interval, where the bit depth determines how detailed the sample will be also known as the dynamic range of the signal (typically 16 bit which means a sample can range from 65,536 amplitude values). Therefore the data will be analyzed for each sound excerpts is essentially a one-dimensional array. Fig. 1 shows the architecture diagram of voice pathology identification of deep neural network.

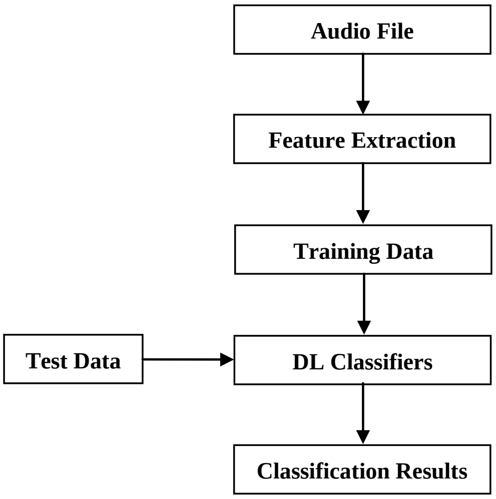

Fig.1 Architecture Diagram

A. Feature Extraction Techniques

Mel Frequency Cepstral Coefficients (MFCCs) is a most common attribute that is mostly implemented in voice identification application. They were defined by Davis and Murmelstein in the 1980's. Before the beginning of MFCC, the linear prediction Coefficients and Linear Spectral Coefficient used as a major feature extraction techniques combined to work with HMM classification models. An audio signal is regularly changing, so audio signal time reduced into 20-40 ms per frame. Next process is to calculate the power spectrum of each signal in use. This is forced by the organ in the ear, which vibrates at different rates depending on the rate of the incoming sounds. Depending on the position in the cochlea that vibrates (which wobbles small hairs), different nerves fire alerts the brains at certain frequencies is present. Periodgram calculate approximately a similar job for identifying which frequencies are present in the frame. The periodgram spectral estimate still contains a lot of irrelevant information. This effect becomes more pronounced as the frequencies increase. Mel filter bank can be examined an indication of how much energy present at near 0 Hertz. Frequency become higher, filters get wider as become less concerned about variations. The Mel scale presents exactly how to space our filter banks and how broad to make them. Take log on the filter bank energies once identified it. This is forced by human hearing don't look loudness on a linear scale. Generally, to increase the actual volume of a resonance need to put 8 times as much energy into it. With effect to large variations in energy may not sound all that different if the sound is loud to begin with. This compression method makes our attributes equivalent more closely what humans actually hear. Last methods to find the DCT of the log filter bank energies. The DCT de-correlates the filter bank energies which means that, diagonal covariance matrices can be applied as a features in a HMM classifier model. Out of 26 coefficients only 12 are taken. So the upper DCT coefficients went for earliest changes in the energies and it turns out that these changes reduce the efficiency of the ASR, hence to result in a little enhancement by dipping them.

The MFCC correlates rate of recurrence, or pitch, of a pure tone to its definite calculated frequency. Humans are much better at perceptive small changes in pitch at low frequencies than they are at elevated frequencies.

IV. BUILDING THE MODEL

Two algorithms are used for building the models.

A. Multi Layer Perceptron Algorithm

A Multi-Layer Perceptron or Multi-Layer Neural Network simulates multiple hidden layers between input and output layer. In Contrast to, single layer Perceptron can learn only linear functions, while a multi-layer Perceptron can learn non-linear functions.

As shown in Fig.2 this model provide an input like X1,X2,...,Xn and outputs f, where f(.) is called as activation function. In order to provide every node with Constant value, Bias is need to implement .The number is given as input to activation function, based on this input it perform some mathematical operation. Many activation function encounter in practice.

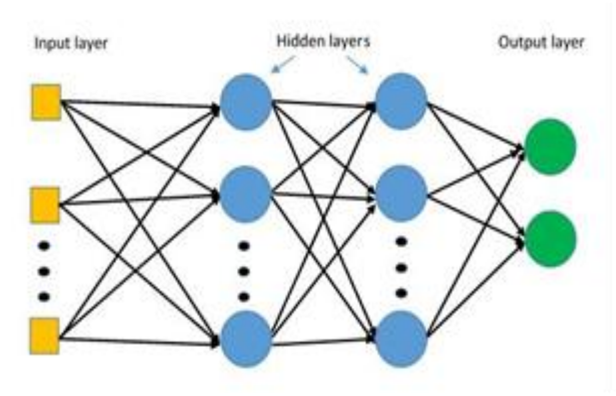

Fig.2 Multi-Layer Perceptron Architecture

B. Convolution Neural Network Architecture

Convolutional Neural Networks is constructed to examine the data through multiple layers of arrays and applied in applications like image recognition or voice recognition. Each parallel layer of a neural network connects some input neurons. Each layer output givens as input to other layers in the CNN. Here weight is associated with each layer. Individuals neurons carry out a move from time to time is called Convolution. The architecture of CNN is shown in Fig. 3.

C. Dataset

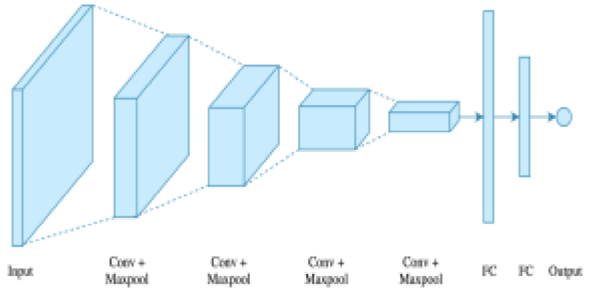

Fig.3 Convolutional Neural Networks Architecture

The original dataset has been collected from special school children. The reading material to identify different disorders are created and given to school children for reading. The voices are then recorded in an audio file. The following Table.1 shows the members of sound excerpts collected from each child to identify various disorders and Table 2 and 3 shows that experimental setup used in our research. We are used our own dataset here. In future, we will collect more data

in both normal and pathological data items.

Table 1. Types of disorder

Disorders	**Files**
Normal	20
Phonological Disorder	18
Speech and Language disorder	33

D, Experimental Setup

Table 2. Experimental setup of MLP

Feature Extraction	**MLP Layers**	**Activation Function**
MFCC	Dense(256)	Tanh, Relu, Sigmoid

Table 3. Experimental setup of CNN

Feature Extraction	**CNN Layers**	**Activation**	**Optimizer**
MFCC	Convolution, filter=16,kernal size=2, max pooling, size=2	Relu	ADAM
	Convolution, filter=32,kernal size=2, max pooling, size=2	Relu	
	Convolution, filter=64,kernal size=2, max pooling, size=2	Relu	ADAM
	Convolution, filter=128,kernal size=2, max pooling, size=2	Relu	ADAM

V. RESULTS AND DISCUSSION

The Objective of feature extraction is to select the audio features in more detail and expressive way such that it is easy to deal with when applying CNN and MLP models. MLP Training is completed with three activations function as discussed before. Tanh is correlated to logistic sigmoid activation function but its efficiency is better. Refer Fig 4 for Accuracy comparison of CNN and MLP with different activation functions

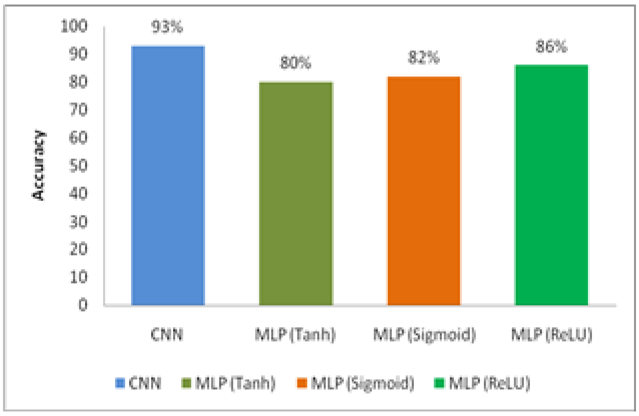

Fig. 4 Accuracy comparison of CNN and MLP with different activation function

Table 4. Overall performance measurement of both CNN and MLP

Techniques/Architecture Used	Performance
Convolutional Neural Network	93%
Multi-layer Perceptron	80%

The Table 4 shows the performance of MLP with different activation functions are measured and compared with CNN.

The MLP performance gets increased with the use of Relu activation function at each layer.

From Table 3, it is obvious that CNN performs well to learn features and predicting the final class.

VI. CONCLUSION

This work is focused on voice pathology detection system was embedded to framework to constantly assess the voice condition of a children.

Deep neural architecture can be applied using CNN and MLP to discriminate between normal and pathology subjects. The feature extraction techniques using MFCC explains that voice detection can be done using these features. The results show the CNN outperforms MLP.

VII. FUTURE WORK

Expansion of this work is the classification of many disorders which describe the childhood onset fluency disorder, social communicational disorder and unspecified communication disorder. One of the future works includes more experiments with different hyper parameters to improve the results and to use other feature extraction techniques for further improvement. Both the Multilayer Feed forward Network with back propagation algorithm and the Recurrent Neural Network can be implemented in future.

VIII. ACKNOWLEDGMENT

This study is supported by the Indian Council of Social Science Research (ICSSR) (F.No IMPRESS/P1565/36/18-19/ICSSR, Dated: 20.03.2019), Govt of India.

REFERENCES

1. Hamzeh Ghasemzadeh, Mehdi Tajik Khass, Meisam Khalil Arjmandi, Mohammad Pooyan, "Detection of vocal disorders based on phase space parameters and Lyapunov spectrum," Biomedical Signal Processing and Control,Volume 22,Pages 135-145.
2. HeatherC. Nardone, MD1; Thomas Recko, BA2; Lin Huang,"A Retrospective review of the progression of pediatric vocal fold nodules," JAMA Otolaryngol Head Neck Surg. 2014;140(3):233-236. doi:10.1001/jamaoto.2013.6378
3. Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk, "Speech Recognition using MFCC", International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) July 28-29, 2012 Pattaya (Thailand)
4. N.J.Nalini, S.Palanivel, "Music emotion recognition: The combined evidence of MFCC and residual phase",Egyptian Informatics Journal (2016) 17,1-10
5. Hyeran Byun,Seong-Whan Lee, "Applications of Support Vector Machines for pattern recongnition: A Survey", SVM 2002: Pattern Recognition with Support Vector Machines pp 213-236
6. Jo, Cheolwoo / Wang, Soo-Geon / Yang, Byung-Gon / Kim, Hyung-Soon / Li, Tao (2004): "Classification of pathological voice including severely noisy cases", In INTERSPEECH-2004, 77-80.
7. Prof. (Dr.) Y. P. Singh, Director, Somany (P.G.) I.T.M., Rewari, " An Approach to Speech Recognition - Challenges & Concept ", International Journal of IT, Engineering and Applied Sciences Research (IJIEASR) Volume 2, No. 12, December 2013
8. Evavan Leer*Robert C.Pfister†XuefuZhou‡, An iOS-based Cepstral Peak Prominence Application: Feasibility for Patient Practice of Resonant Voice, Journal of Voice, Volume 31, Issue 1, January 2017, Pages 131.e9-131.e16
9. Laura Verde ; Giuseppe De Pietro ; Giovanna Sannino, Voice Disorder Identification by Using Machine Learning Techniques, IEEE Access (Volume: 6)
10. Shi, S.; Wang, Q.; Xu, P.; Chu, X.:" Benchmarking state-of-the-art deep learning software tools" , In Cloud Computing and Big Data (CCBD), 2016 7th IEEE Int. Conf., pp. 99–104, IEEE, 2016.
11. Zhenzhou Wu1 , Sunil Sivadas2 , Yong Kiam Tan1 , Ma Bin2 , Rick Siow Mong Goh, "Multi-Modal Hybrid Deep Neural Network for Speech Enhancement"
12. Björn W.Schuller, Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning, Health Informatics and Translational Data Analytics
13. Aditya Khamparia ; Deepak Gupta ; Nhu Gia Nguyen, Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network, New Trends in Brain Signal Processing and Analysis, 2169-3536

AUTHORS PROFILE

Dr. C. S. Kanimozhi Selvi is a professor in the Department of Computer Science and Engineering of Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India. She holds a Ph.D. in Computer Science (2011) from Anna University, Chennai. She is at the teaching profession for more than 20 years and more than 10 years research experience. Her areas of academic interest include data mining, Machine Learning and Deep Learning. She has published more than 40 articles in international journals and more than 30 papers in international and national conferences.

Mr.M.Balaji Prasath received the B.Tech Degree in IT from Sona College of Technology, Salem, Tamil Nadu in 2007 and M.E in CSE from Sona College of Technology, Salem, Tamil Nadu. He is working as Research Assistant in Kongu Engineering College, Erode, Tamil Nadu. His research interest includes Machine Learning, Data Mining, and Computer Networks.

Miss.T.Sathiyawathi is a student member in the Department of Computer Science and Engineering of Kongu Engineering College, Perundurai, Erode. She received her Bachelor Degree B.E., in Computer Science and Engineering from Kongunadu College of Engineering. Her area of interest is Deep learning.