An evaluation of deep learning approaches for detection of voice disorders.

Chinchu M S¹, Ms. Dr. B. Kirubagari ²

¹ Assistant Professor at SJCET Palai & Research Scholar, Computer science and Engineering Department, India
² Associate Professor & Research Guide, Computer science and Engineering Department, India.
chinchums18@gmail.com, kirubacdm@gmail.com

Abstract. The human voice manufacturing system is a complicated natural device capable of modulating pitch and loudness Human sound frequency particularly. The part in which the folded is the primary source underlying internal and/or external factors often destroys vocal folds justification. Some changes consequences are reflected in functioning emotional state soul. Therefore it essential to identify variations at an early stage gives the patient an endangerment overcome any impact modernize their quality of life will-less detection disorders using depth study methods plays an important role as has been proven to facilitate the process. Many researchers have explored technologies for streamlined that can help clinics diagnose noise paper we present the conducted research activities.

1. Introduction
Voice order is a health issue that is neglected often by the people as the people neglected the superintendency of the voice or not desired to meet the professional doctor. Previous researches has developed voice disorder detection approaches using the sensory features and machine learning from the voice recordings rather than using the professional medical devices for the detection. There are two classes of voice recordings, like Pathological and Normal. These classes are explored by the increasing popularity of full-length learning, deep learning and transfer learning. The voice disorders are recognized well using the unstipulated viability of full-length learning and deep learning. The voice disorders are recognized well using the unstipulated viability of full-length learning and deep learning. Moreover, in the existing datasets, the voice disorder detection like the lack of generality and size of the dataset are the major drawbacks.

The speech is defined as the fundamental instinct of the human and voice in its subsystem. The study of vocalization is termed as, vocology [1]. The interaction of sound product of the pulmonary air pulses interacting with the trachea is the Normal sound. The Normal sound produced the periodic and/or aperiodic sounds by snatching the original vocal folds. Some of the noise disorders include the dysphonia and aphonia caused by multiple wiseacre vocalizations. The dysphonia is the partial sound loss, whereas the aphonia is the complete sound loss. The disorder in the voice occurs due to the gender, age and social group that has variations in the volume, quality, and tone [2]. The impact of the disorders in the malignant speech is not life-threatening. However, the speech disorders that are professional, social, and personal aspects of the liaison are not affected by the impact of the untreated speech disorders [3]. The increased public sensation of the voice care needs is caused due to the changes in the professional and lifestyle needs. Now-a-days, the citizen consults the voice therapists and laryngologists for maintaining the voice health. However, the examinations, like stroboscopy, laryngoscopy or videos are expensive and difficult. Speech processing is a developing research zone that focuses on aspects of the speech processing from full-length extraction to visualization support systems.
2. Voice Disorder
The disorder in communication science are categorized into five types, such as language disorders, speech disorders, cognitive and social communication disorders and swallowing disorders [5]. According to the Classification Manual for Voice Disorders, the hypo-functional and hyper-functional vocal fold pathologies are classified as,

- Injury or Trauma-based,
- Inflammatory conditions, such as acute laryngitis,
- Structural lesions, such as vocal polyps and vocal nodules,
- No laryngeal aero digestive disorders, consisting of bronchitis and reflux disorder among others.
- Neurological conditions, like abductor palsy, adductor palsy and spasmodic dysphonia,
- Systemic conditions, such as, hypothyroidism and hyperthyroidism,
- Psychological disorders and psychiatric disorders that are functional and mutism
- Undiagnosed and otherwise not specified.
- Other disorders, such as muscle tension dysphonia [5]

Mass-based pathologies of various vocal folds are very popular. The vocal polyp and vocal nodule occurred due to the external influence and persistent tissue inflammation [6]. In these circumstances, the voice production is conceptually rough and not economical, whereas the vocal fold closure is incomplete. Even though the no vocal folds in the nonphotromatic voice disorders, the quality of the voice is decreased along with increased laryngeal tension and noise fatigue. The nonphotromatic voice disorder includes functional voice disorder and muscle tension dysphonia. For evaluating the voice, the Multi-parametric assessment protocol is taken into consideration [2], [7]. The laryngeal imaging from the laryngoscopy and stroboscopic, patient inter- view, standardized psychoacoustic methods for perceptual analysis, evaluation of subjective voice, assessment of aerodynamic and acoustical analysis is essential for the structured approach. In order to distinguish the normal voice from the dysphonia and/or aphony, the recent technologies are developed by the scientists for developing tools for acoustical analysis. The tools should be developed to solutions in financial that are reliable in dynamic clinical settings and sensitive in detecting the changes in the sound. If enough research is done on acoustic analysis, the result is an automatic discovery.

Recently, numerous solutions are developed in the pathological speech detection field [21], [22]. For instance, Al-Nasheri et al.[23] designed a extraction method for the classification and detection of voice pathology using entropy and autocorrelation by determining the variegated frequency bands. According to the study, the vocal polyps, vocal cysts and vocal paralysis are caused due to the impairment of voice. For the classification and detection, the frequency band is considered between 1000 Hz and 8000 Hz. The support vector machine is used as a classifier and each voice sample consists of sustained vowel /a/. The accuracy obtained for the detection of specimen are, 92.79%, 99.69%, and 99.79% for Saarbrucken Voice Database (SVD), Massachusetts Eye and Ear Infirmary (MEEI) and Arabic Voice Pathology Database (AVPD), respectively.

3. Voice Disorder Database
The dataset consists of the recordings of pathological and normal sounds. The audio recording has continuous conversation and regular vowel pronunciation. From the survey, it is observed that most of the researchers use the standard database, like SVD, MEEI and AVPD. The information from the three databases described in Nasheri et al.[8] are considered. Other researchers used the dataset from the private dataset in association with the local hospitals are considered. In the dataset, the audio recording of pathological and normal sounds is taken into consideration. The recordings in the dataset
comprises of continuous conversation or regular vowel pronunciation. The standard databases, like Saarbrucken Voice Database (SVD), Massachusetts Eye and Ear Infirmary (MEEI), and Arabic Voice Pathology Database (AVPD). The public misogynist Saarbrucken Voice Database is also used by most researchers [20]. In public misogynist Saarbrucken Voice Database, the voice recordings comprises of the samples of the subject with the recordings of vowels, /a/, /i/, and /u/ that are produced at high, normal, low and low-high-low pitches. The recording length ranges from 1 to 2 seconds for sustained vowels.

### 3.1. Feature Extraction

The feature extraction is the process of extracting the relevant feature from the test set. The feature extraction process is carried on for reducing the feature vector size with respect to the dimension of the feature extraction. The redundant features are removed in the feature extraction process. This process reduced the computation time and improved the accuracy for helping the learning algorithm. The complexities can be further determined using other machine learning methods. In the NLP, repetitive neural networks are applied frequently when compared with the other models. The data is targeted for particular time range in the RNN. Similar features are extracted using multiple extraction methods. For the voice disorder detection system, the first step is the feature extraction. The frequently used methods for the feature extraction and acoustic analysis are discussed below.

#### 3.1.1. Mel-Frequency Cepstral Coefficient (MFCC)

The MFCC is a standardized process for extracting features that utilize knowledge of the human auditory system. A popular audio feature extraction method is the Mel-frequency septic coefficient (MFCC) with 39 features. The number of features is small enough to motivate us to understand the audio information. 12 parameters are related to the magnitude of the frequencies. It provides us enough frequency channels to analyze the audio. The process of extracting the MFCC features from the single frames [12], [13] are given as,

a) Computing the coefficients of discrete Fourier transform.
b) Using Mel spaced triangular filter for filtering.
c) Computing the energies of sub-band.
d) Computing the coefficients of discrete cosine transform.
3.1.2. **Neural Networks.**

The neural network is modelled on the basis of human brain for recognizing the patterns. The sensory data is interpreted with the form of labelling, machine perception, or raw input clustering. All the real-world data, like sound, images, time series or text are translated by the patterns that are contained in vectors and numeric. The features from the other algorithms are extracted for the classification and clustering process. The neural networks are used as a layer helps in categorizing and clustering the features that are stored and managed. For instance, the inputs are used for grouping the unlabelled data depending upon the similarities and when there is a labelled dataset for training, they classify the data. The correlation between the person name and image pixel are established using the deep learning, which is known as a static forecast. The exposure to the data is done adequately with the help of same token for establishing the relationship within the future events and current events in the deep learning study. There is no need of worrying for the study time of the deep learning. Then, the number is predicted and string of numbers is read for the given time range in the deep learning study. The unstructured data are applied with the labelling data from the trained in-depth learning network for providing access to the input for improving the performance. The accuracy of this method depends on the number of data that can be trained by the network. In the in-depth learning network, the raw data is provided in the form of image and the individual is represented with 90 percent chance from the input data.

**Deep Neural Networks.** The Deep learning network uses the single-hidden-layer neural networks for separating from their depth. The data is passed through a multistep process for the recognition of pattern. The layers in in-depth learning networks are trained with specific features depending on the output of the preceding layer. As you progress further into the neural network, the features that your nodes can distinguish are more complex because they combine and reassemble features from the
previous layer. The automatic feature extraction is performed in the deep-learning without the intervention of humans [7, 8]. In the deep-learning network, more access is provided to the input by training the labeled data and applying to the unstructured data. The accuracy of the deep-learning network depends on the data for training the net. When compared to the previous algorithms, the ability of the deep learning to learn from the unlabeled data provides a huge advantage. The likelihood is assigned to the specified label or outcome by the deep learning networks.

3.1.3. Convolutional Neural Network (CNN)

CNNs are the neural networks used for object identification in scenes, similar cluster images and sorting the images based on scene. For instance, CNN are used for identifying the street signs, faces, individuals, platypus, tumours and other aspects of data. The images are considered as volumes in the conversion chains. The convolutional chains are three-dimensional objects rather than the flat canvases that measured the height and width. All the colours are combined together to form the colour spectrum in the digital colour images with red-blue-green (RGB) encoding. The images of different colour levels are stacked together on the top of another using the conventional network. Hence, the standard colour image is considered as a rectangular box in the convolutional network in which the height and width are measured with the help of the number of pixels in the dimension, whereas the depth is indicated by the three layers that represents the letter in RGB. The channels are considered as the depth layer.

Huiyiet al. [8] developed a voice detection approach using the CNN. In this approach, the pathological speech recordings are given as the network input. The weight in the CNN is trained with the help of Convulsive Deep Confidence Network (CDBN). The input data structure is determined using the statistical methods in the generative model. The weights are improved by training the CNN using the supervised back-propagation learning algorithm. Better results are achieved during the classification through this approach.

Musaed alhussein and Ghulam Muhammad et al. [9] designed a Mobile Healthcare Framework for voice detection using deep learning. The sounds in the detection system are recorded with the smart mobile devices. The voice signals are pre-processed and provided to the CNN. The CNN models are used along with the transfer learning method, especially the VGG-16. The detection accuracy is up to 97.5% in the voice pathology detection process.

Alice Rueda and Sridhar Krishnan et al. [10] designed a voice detection method using the Fourier-based synchrosqueezing transform in the CNN. One of the major challenges in the study of dysphonia voice is the small dataset. The in-depth learning methods that does not have fitting are more difficult to apply. Large amount of training data is required by the CNN and the sound is limited for the data enhancement methods. The data size is enhanced using the data augmentation technique, such as Fourier-based Synchronous Caching Transformation (FSST). Besides increasing the size of the data, FSST method provided better study with the CNN than the Short-Time Fourier Transform (STFT). Different vowels are constructed in the CNN using the dynamic stopping algorithm. For the first selected small dataset, the results are trained by improving the CNN in advance. The fitting is bypassed using larger dataset by improving the performance using the increased number of convex layers. More reliable and meaningful results are generated using the larger database.

Huiyi Wu and John Soraghanet al. [11] developed a Voice Detection approach using CNN. The sound is classified as, healthy or pathological by extracting the sound features through acoustic analysis with the help of signal processing tools. The automatic feature are extracted and classified into normal and random sounds by the spectrogram specific voice recording in the database. The accuracy obtained for the dataset containing the 482 organic dysphonia speech files and 482 normal dysphonia speech files are set as, 88.5% for training, 66.2% for evaluation, and 77.0% for test data.
3.1.4. Recurrent Neural Network (RNN)

RNN are the type of NN algorithms used for continuous processing of the data like time range data, voice or written natural language. A version of RNN is used by the Deep Mind for playing the video games along with the autonomous agents. At each subsequent step, the RNN returned the output to the RNN from \( n-1 \) step without affecting the \( n \) step with the help of feedback loop. When compared to the traditional computers, the RNN allows continuous and parallel calculations. The RNN is in the shape of human brain that consists of large network of neurons connected together to translate the input stream into a series of output motor [14]. One input and output is received and generated at a given time using the Feed forwarded networks. Same type of obstruction is not faced by the RNN. When submitted to the RNN, the image data is appeared as a sequence and they are not continuous. For example, if the image is a handwritten word then, the cursive image is transformed into a letter. The word ending is predicted by figuring out the beginning of the word. Thus, the image portion is considered as a series of letters. The problems in the traditional RNNs are overcome by the feedback network, known as Long Short-Term Memory (LSTM). The LSTM solved the problems occurred in the previous tasks.

- In noisy input sequences, the extended patterns are temporally recognized.
- The events in the noisy input streams are separated widely by recognizing the temporal order.
- The temporal distance between the events are conveyed by extracting the information.
- The non-smooth and smooth periodic trajectories, precisely timed rhythms are generated in stable.
- The high-precision real numbers are stored robustly across the extended intervals of time.

3.15. LSTM Networks

The long-term dependency is dealt with the help of RNN. The Long-term memory networks are a type of RNN that is commonly known as, LSTM. The long-term dependence is avoided by designing the LSTM clearly. One of the default behaviors of the LSTM is the storage of information over long intervals of time. There are more repetitive modules in the neural networks.

Vibhuti Gupta et al. [16] developed a voice disorder model using LSTM. The performance of the voice disorder method is determined by considering the original 400 testing samples without labels. The LSTM model is used for the classification by extracting the features using different feature extraction methods.

Danish Raza Rizvi et al. [17] designed a Deep Learning Model for detecting the voice disorder. The Parkinson's disease (PD) is predicted by determining the LSTM and DNN. In this method, the accuracy obtained is 97.12% and 99.03% for varying sizes. Junli Gao et al. [18] designed an Effective LSTM Repeat Network for voice disorder method. The training effect is enhanced using the LSTM network along with FL for eliminating the impact of the standard ECG beat data on training the model. This method provided accuracy, recall, accuracy, specificity, and F1 score of 99.26%, 99.26%, 99.30%, 99.14%, and 99.27%, respectively.

4. Conclusion and Future Works.

In this research, we have analysed the major research activities related to detection approach of voice disorders system. This paper is based on the organization of the contents, sound distortions, feature extraction, and DL techniques used. The theoretical aspects of voice disorders, feature extraction methods, and DL methods, and performance of the significant research activities in this field are reviewed. From the standard database, the subjects of certain diseases are selected for performing the experimental activities. In most papers, sounds are classified as normal and pathological. However, PD and Alzheimer's disease are specifically targeted in few studies. The research activities in this field have great potential as they apply to the community. With this in mind, we need to develop more
standard databases and discover new features. The standard database conducts future research activity experiments by selecting a few randomly selected voice disorders from SVD. It was decided to classify sounds as normal and pathological using DL techniques.

References

[1] A Survey on Machine Learning Approaches for Automatic Detection of Voice Disorders. Journal of Voice, Vol. &&, No. &&, 2018.(Placeholder1).
[2] Development of the Arabic Voice Pathology Database and Its Evaluation by Using Speech Features and Machine Learning Algorithms.
[3] Towards Robust Voice Pathology Detection-Investigation of supervised deep learning, gradient boosting, and anomaly detection approaches across four databases.
[4] Fundamentals of Speech Recognition. Englewood Cliffs: PTR Prentice Hall; 1993
[5] High-Resolution Motor State Detection in Parkinson’s disease using convolutional neural network-2020 Franz M. J. Pfister,
[6] Sung hwon Yoo, Isha Gujrathi, Masoom A. Haider & Farzad Khalvati, Prostate cancer Detection using deep convolutional neural networks -2019.
[7] Pavol Harar*, Jesus B. Alonso-Hernandez†, Jiri Mekysk Voice Pathology Detection Using Deep Learning on Mobile Healthcare Framework, volume 6, 2018.
[8] Mazin Abed Mohammed 1,*OrcID,Karrar Hameed Abdulkareem 2OrcID, Salama A. Mostafa Augmenting dysphonia voice using Fourier-based synchro squeezing transform for a CNN classifier, 978-1-5386-4658-8/18/$31.00 ©2019 IEEE.
[9] Huiyi Wu, John Soraghan Senior Member, IEEE, Anja Lowit, Gaetano Di Caterina Convolutional Neural Networks for Pathological Voice Detection-2018
[10] Convolutional Neural Networks for Raw Speech Recognition, Convolutional Neural Networks for Raw Speech Recognition http://dx.doi.org/10.5772/intechopen.80026.
[11] Convolutional Neural Networks for Pathological Voice Detection, Conference: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).
[12] Speech Recognition Using Convolutional Neural Networks. International Journal of Engineering & Technology, 7 (4.6) (2018) 133-137.
[13] Mahmoud Keshavarzi 1, Tobias Goehring 2, Justin Zakis, Use of a Deep Recurrent Neural Network to Reduce Wind Noise: Effects on Judged Speech Intelligibility and Sound Quality-2018.
[14] Vibhuti Gupta, Voice Disorder Detection Using Long Short Term Memory (LSTM) Model-2018.
[15] Danish Raza Rizvi1, Iqra Nissar2, Sarfaraz Masood*3, Muntaz Ahmed4, Faiyaz Ahmad5 An LSTM based Deep learning model for voice-based detection of Parkinson’s disease-2020.
[16] An Effective LSTM Recurrent Network to Detect Arrhythmia on Imbalanced ECG Dataset-2019
[17] Development of Output Correction Methodology for Long Short Term Memory-Based Speech Recognition-2019
[18] S. Girirajan, R. Sangeetha, T. Preethi, A. Chinnappa, Automatic Speech Recognition with Stuttering Speech Removal using Long Short-Term Memory (LSTM) -2020

[19] Barry, B. Saarbruecken Voice Database. Institute of Phonetics; Saarland University. Available online: http://stimmdb.coli.uni-saarland.de/ (accessed on 23 February 2017).

[20] Gupta, R.; Chaspari, T.; Kim, J.; Kumar, N.; Bone, D.; Narayanan, S. Pathological speech processing:
   a. State-of-the-art, current challenges, and future directions. In Proceedings of the 2016 IEEE International
   b. Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March

[21] Danubianu, M.; Pentiuc, S.G.; Schipor, O.A.; Tobolcea, I. Advanced Information Technology-support of improved personalized therapy of speech disorders. Int. J. Comput. Commun. Control 2010, 5, 684–692.2016; pp. 6470–6474

[22] Al-Nasheri, A.; Muhammad, G.; Alsulaiman, M.; Ali, Z.; Malik, K.H.; Mesallam, T.A.; Ibrahim, M.F. Voice Pathology Detection and Classification using Auto-correlation and entropy features in Different Frequency Regions. IEEE Access 2017, 6, 6961–6974