An Evolving Model of Voice Disorder Detection using Deep Belief Network

P. Kokila[1*], Dr. G. M. Nasira[2]
[1]*Research Scholar, Department of Computer Science,*
Chikkanna government Arts college, Tirupur-641602, Tamil Nadu, India
[2]*Professor & Head, Department of Computer Applications,*
Chikkanna Government Arts College, Tirupur-641602, Tamil Nadu, India.

Abstract - *In recent years, automatic diagnose of larynx pathological voice disorders are a challenging task in the medical filed. The researchers started focusing on working with voice signals to discover voice disorder related diseases. Machine learning plays a vital role in automatic detection of voice disorder using spectral information of recorded voice. Among several approaches deep playing has been in a prominent place for achieving significant results in the voice recognition field, where there has been less research work in the field of pathological voice detection. This paper introduces the deep belief network for discovering healthy and unhealthy voice detection. The stack of Restricted Boltzmann Machine is used to pretrain the deep neural networks. Simulation analysis is done to prove the proficiency of the deep belief network-based voice disorder detection using the real data from the Saarbrucken Voice database.*

Keywords — *Voice disorder, deep learning, deep belief network, Restricted Boltzmann Machine, pathological voice, machine learning*
.

I. INTRODUCTION

Voice pathologies affect the larynx and result in irregular vibrations of the vocal folds. Poor voice can impact on individual's ability to communicate both socially as well as in the work place, thus reducing quality of life, and it has a significant impact on economy considering the costs of medical diagnosis and treatment[1]. Traditional diagnostic method of voice pathologies relies on clinician's experiences and on expensive devices such as laryngoscope, endoscope etc. However, computer-aided medical systems for diagnosis of voice pathologies have been popular due to major advance in signal processing techniques. These complementary tools are usually non-invasive and nonsubjective, which generally are an advantage in medical field. A lot of research related to automatic detection of voice pathologies has been carried out in the past few decades. In this context, features are extracted from the speech recordings and they are then processed by classifiers to distinguish normal voice instances from pathological voice recordings. These features are mainly derived from two research fields. One is from speech recognition applications, with signal processing tools used to automatically detect features such as MelFrequency cepstral coefficients (MFCC), linear prediction cepstral coefficients (LPCC) and energy and entropy of discrete wavelet packets[2-4]. Other features come from voice quality measurement according to physiological and etiological research.

While pitch, jitter and shimmer are used to detect the roughness of the speech, other characteristics such as harmonicto-noise ratio (HNR), normalized noise energy (NNE), glottalto-noise ratio (GNR) and cepstral peak prominence (CPP) represent the breathiness of the speech[5]. Most of the research works use the Massachusetts Eye and Ear Infirmary (MEEI) database. However, healthy voice recordings and pathological voice recordings in this database are recorded in two different environments[6], which make it hard to distinguish whether it is discriminating environments or voice features. The Saarbruecken Voice Database is a downloadable database with all recordings sampled at 50 kHz and with 16-bit resolution. This database is relatively new so that little research has been carried out through it. However, the audio samples are recorded in the same environment so that it is an ideal database for this work.

Related Work

In this section few of the existing works on pathological voice disorder detection is discussed.

In [9] the authors explored the information collected form acoustic and modulation frequency representation are used to detect and classify the discrimination of voice disorders. The input is converted to a low dimensional domain by adapting higher order singular value decomposition. Using Mutual Information, the feature selection is achieved.

In [10] the authors have developed a vocal fold paralysis recognition method using amplitude modulation and features are extracted using MFCC integrated with GMM. The equal error rate is reduced in this method.

Markaki et al. [11] explored the information provided by a joint acoustic and modulation frequency representation, referred to as modulation spectrum, for detection and discrimination of voice disorders. The initial representation is first transformed to a lower dimensional domain using higher order singular value decomposition

(HOSVD). For voice pathology detection an accuracy of 94.1% was achieved using SVM as classifier

In Paneket al.[12], a vector made up of 28 acoustic parameters is evaluated using Principal Component Analysis (PCA), kernel principal component analysis (kPCA) and an auto-associative neural network (NLPCA) in four kinds of pathology detection (hyperfunctional dysphonia, functional dysphonia, laryngitis, vocal cord paralysis) using the /a/, /i/ and /u/ vowels, spoken at a high, low and normal pitch. The results show a best efficiency level of around 100%.

Al-Nasheriet al.[13] investigated different frequency bands using correlation functions. The authors extracted maximum peak values and their corresponding lag values from each frame of a voiced signal by using correlation functions as features to detect and classify pathological samples. Three different databases were used, Arabic Voice Pathology Database (AVPD), Saarbruecken Voice Database (SVD) and Massachusetts Eye and Ear Infirmary (MEEI). A Support Vector Machine was used as classifier. For detection of pathology an accuracy of 99,8%, 90.9% and 91.1% was achieved for the three databases respectively. In classification of the pathology task an accuracy of 99.2%, 98,9% and 95.1%, respectively, was achieved for the three databases

Hugo Cordeiro1 [14] presented a set of experiments to identify the best set of features from the vocal tract (MFCC, Line Spectral Frequencies (LSF), Mel-Line Spectral Frequencies (MLSF) and first peak of the spectral envelop) and the best classifiers amongst SVM and Gaussian Mixture Models (GMM) for the identification of pathologic voices. He achieved an accuracy of 84,4% for the identification between 3 groups (healthy subjects, subjects with physiological larynx pathologies - vocal fold nodules and edemas, and subjects with neurological larynx pathologies - unilateral vocal fold paralysis). He also used Regression Trees to the pathological voice recognition based on formant analysis and harmonic-to-noise ratio with 95% of recognition rate.

In this paper the deep learning model is used for voice disorder detection which involves in strengthening the process of classification of pathological and healthy voice.

.

Methodology of Deep Belief Network based Voice Pathology detection

In this paper, a novel deep belief network is used to automatically discriminate the pathological voice and healthy voice. Deep belief network structure is utilized in this work to analyze the spectrograms of voice recording. Figure 1 shows the block diagram of proposed pathologicalvoice detection system. First, pre-processing steps, such asresampling, reshaping techniques, are applied to the speechrecordings. Meyer Wavelet transform (STFT) technique isthen applied to compute the spectrograms of the speechrecordings as the input to the DBN system. Weights in the DBN system is pre-trained using RBM and fine-tuned with backpropagationmethod. The trained SBN system is capable ofextracting features automatically and classifying audio samples

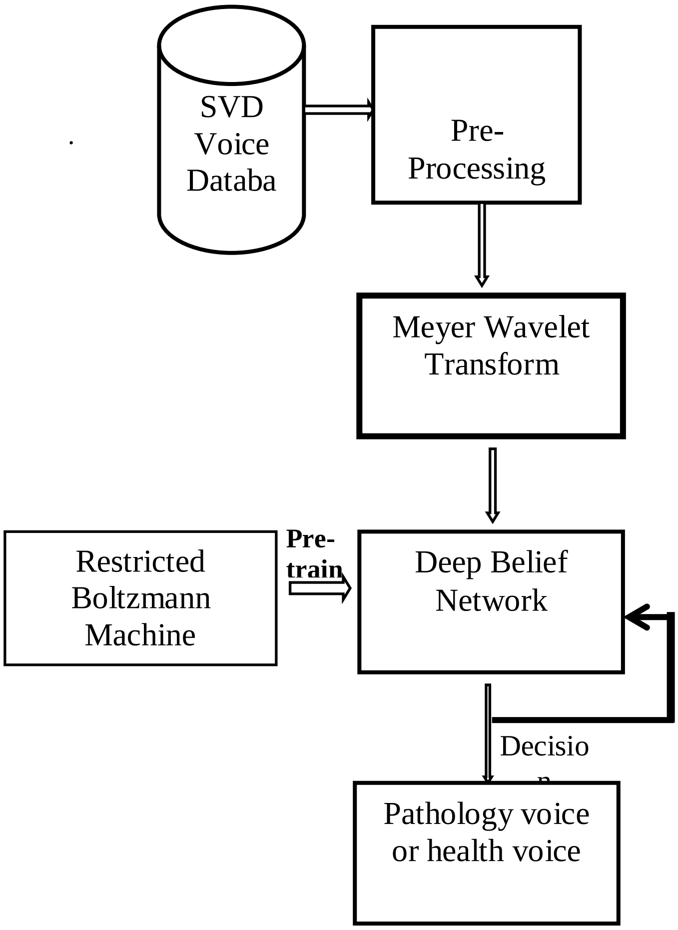


Figure 1 Block Diagram of DBN based Voice disorder Detection

Deep Belief Network is a class of deep neural network which comprises of multiple layer of graphical model having both directed and undirected edges. It is composed of multiple layers of hidden units, where each layer is connected with each other but units are not. The two significant caveats of Deep Belief Networks are:

- Belief Network
- Restricted Boltzmann machine

Belief Network

It consists of stochastic binary unit layers where each connected layer has some weight. The stochastic binary units in belief networks have a state of 0 or 1 and the probability of becoming 1 is determined by a *bias and weighted input from other units. A belief net is a directed acyclic graph which is composed of stochastic variables. It helps in solving two issues they are by inferring states of the unobserved variables and adjusting interaction among variables to enhance the network to produce more likely output data. The general structure of Belief network is shown in the figure 2*

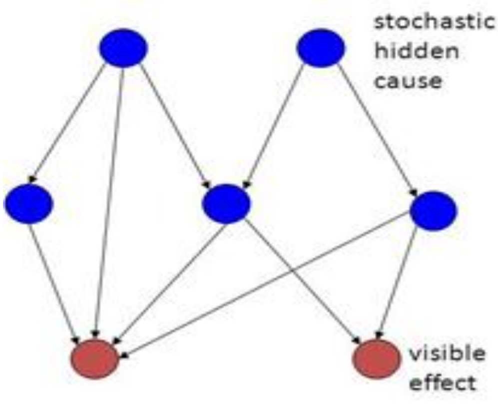


Figure 2: General Structure of Belief Network

Restricted Boltzmann Machines

Restricted Boltzmann Machines (RBM) [7] is a two layered bipartite graph. It comprises two different units namely visible units and hidden units. Each visible unit is connected to all hidden units through a weight matrix

bipartite graph with two layers. It consists of visible units $\{0,1\}$ D v $\in$ and hidden units $\{0,1\}$ P h $\in$, where every visible unit is connected to all hidden units by a weight matrix, as shown in Fig.3 a and b, while the units do not connect with each other within the same layer

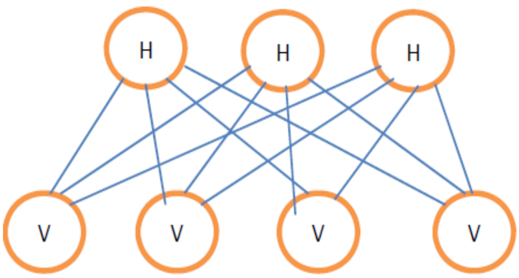


Figure 3: simple RBM layers

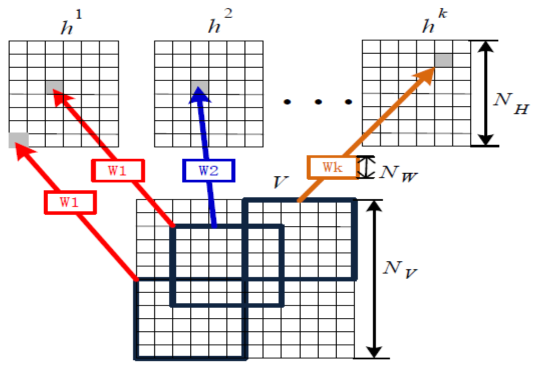


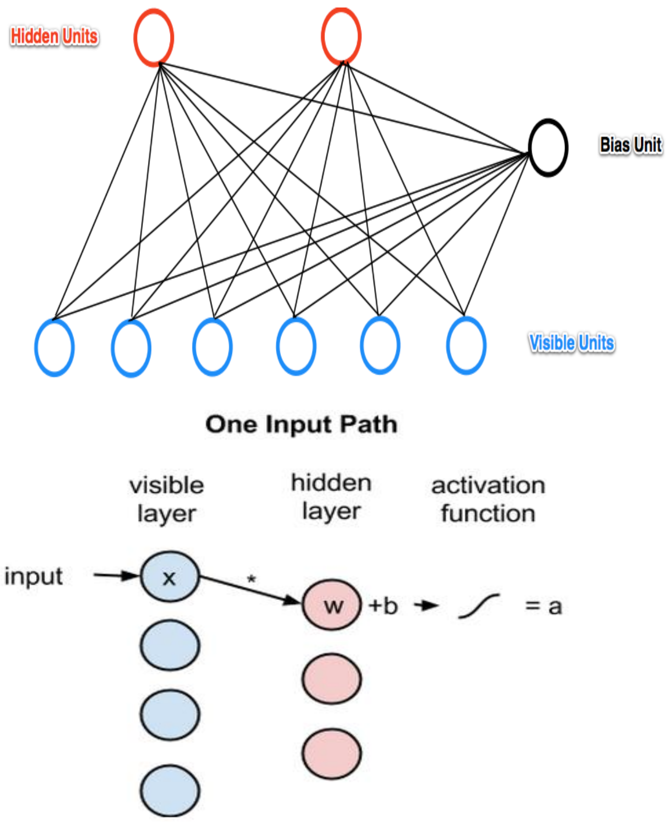
Figure 3: simple RBM layers

The conventional Restricted Boltzmann machine consisting of:

- One layer of **visible units**
- One layer of **hidden units** and
- A bias unit

Each visible unit is connected to all the hidden units, and the bias unit is connected to all the visible units and all the hidden units. In RBM no visible-visible units and hidden-hidden units are connected.

Figure 4: function of one input path of RBM



For the figure 4 it is illustration it is depicted in the figure that how single pixel value, x, through the two-layer net. At node 1 of the hidden layer, x is multiplied by a weight and added to a so-called bias [8]. The result of those two operations is fed into an activation function, which produces the node's output, or the strength of the signal passing through it, given input x.

Act_fun((weight w * input x) + bias b) = output a (1)

Input Data to Deep Belief Network system

A DBN contains "feature extractors" which are commonlyapplied to feature maps. Therefore, speech recordings aretransformed from one-dimensional signals to two-dimensionalspectrograms.

Database

This work uses the Saarbruecken voice database which wasrecorded by the Institute of Phonetics of Saarland University inGermany [15]. This database contains 71 different pathologies withspeech recordings from over 2000 individuals. Each participantfile contains recordings of sustained vowels /a/, /i/ and /u/ inneutral, low, high and low-high-low intonations

Pre-processing and organization of input data

First, the original speech is resampled at 25 kHz in the preprocessingstep. The aim of this step is to reduce the amount ofdata in feature map to boost the training process. Furthermore, Meyer Wavelet Transform is applied to transform the time-domain signal intospectral-domain signal. In this step, each file is divided into 10ms

Hamming window segments, with 50% overlap betweenconsecutive windows. Finally, the spectrogram is reshaped toas common size of 60*155 points to remove parts which containno information. In this case, unwanted noise is dismissed andessential features are preserved. The comparison of inputfeature maps between normal voice and pathological voice isshown in Figure 1.

Experimental Setup

The framework for the training process was developed inPython using Tensorflow. Training data is divided as 256samples in each mini-batch, and is trained with GPU NvidiaGTX1070 for higher speed. DBN sparsity is set as 0.6 and weights pre-trained in thefirst two DBN-RBM layers are set as initialization of DBN. We usesustained vowel /a/ at neutral pitch of each individual, of which482 are healthy and 482 are diagnosed with pathologies We usesustained vowel /a/ at neutral pitch of each individual, of which482 are healthy and 482 are diagnosed with pathologies

Performance Analysis Results

The tables 1 and 2 shows the classification result of different metrics such as Precision, Recall, F-measure, Specificity and Accuracy

Precision:

It defines what proportion of patients that the model diagnosed as have pathology, actually had voice pathology. The predicted positives and the people actually have a voice pathology are known as true positive

$$\text{Precision} = \frac{\text{True positive (actual positives)}}{\text{True positive} + \text{false positive (predicted Positives)}}$$

Recall:

It defines at what proportion of patients that actually had voice pathology are diagnosed by the models as have pathology. The actual positives and the people diagnosed by the model have a pathology in voice are Ture Positive. It is also referred as sensitivity.

$$\text{Recall} = \frac{\text{True positive (actual positives)}}{\text{True positive} + \text{false Negative (Actual number of patients having voice pathology)}}$$

F-measure: It defines a score of combining both Precision and Recall

$$\text{F-Measure} = \frac{2 * Precision * recall}{(Precision + Recall)}$$

Accuracy: In voice disorder classification, the number of correct predictions produce by the model over all kind's prediction models known as accuracy

$$\text{Accuracy} = \frac{\text{True positive (actual positives)}}{\text{True positive} + \text{False Positive} + \text{True Negative}}$$

Specificity: It is defined as proportion of patient that are not have pathology, were predicted by the model as healthy. The actual negatives and the people diagnosed by the model as not having pathological voice are True Negative.

$$\text{Specificity} = \frac{\text{True Negative (actual negatives)}}{\text{False Positive} + \text{TureNegative (Actual number of patients with healthy voice)}}$$

Where

- True Negative: Healthy voice recordings arecorrectly detected
- True Positive: Pathological voice recordings are correctly detected
- FalseNegative (FN): Pathological voice recordings aredetected wrong
- False Positive (FP): Healthyvoice recordings are detected wrong.

Table 1: Performance Analysis based on Precision, Recall and F-measure of three different classification models

True:

	Precision	Recall	F-measure
ANN	0.68	0.75	0.71
SVM	0.65	0.72	0.68
DBN	0.94	0.9	0.92

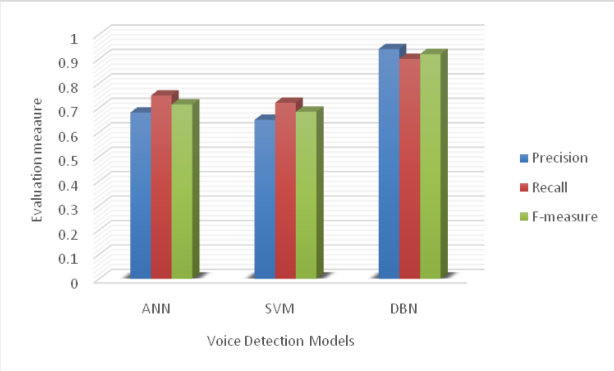


Figure :Performance Analysis based on Precision, Recall and F-measure of three different classification models

The table and the figure show the performance comparison of Artificial Neural Network, Support Vector Machine and proposed Deep Belief Network.

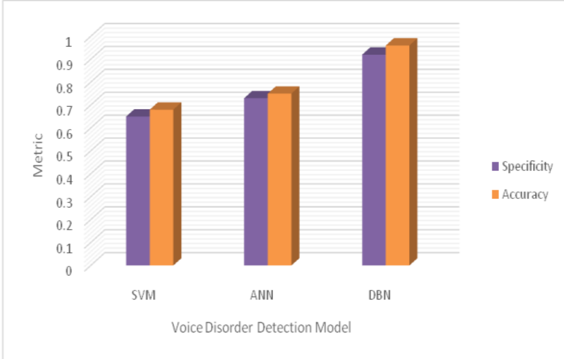
	Specificity	**Accuracy**
SVM	**0.65**	**0.68**
ANN	**0.73**	**0.75**
DBN	**0.92**	**0.96**

The results proved that the performance of DBN produces better result because the nature of its deep learning behaviour and extracting optimal feature vectors which mainly contributes more in higher rate of precision, recall and f-measure while comparing the existing models SVM and ANN. The existing models fails to discover significant independent features involved in voice pathology detection.

Table 2: Performance Analysis based on Specificity and Accuracy of three different classification models

Figure Performance Analysis based on Specificity and Accuracy of three different classification models

From the table and the figure, it is observed that based on the measures of specificity and Accuracy the performance of the Deep belief network provides more promising result compared to the other two models SVM and ANN. This is because the deep learning model consist of stack of Restricted Boltzmann machine which is involved in learning process, additionally the fully connected layer is used as backpropagation to classify the voice as pathological or healthy.



CONCLUSIONS

in this work a novel deep learning model is developed for pathological voice detection. the deep belief network extract the feature vector using stack of restricted boltzmann machine. rbm extracts the features of spectrogram of voice recordings and diagnose the voice disorders. deep belief network assist in initializing weights on the hidden nodes of the entire network and thus it makes the classification model more robust. the simulation results of the dbn is compared with other existing models namely svm and ann. the ability to handle the voluminous feature space of voice signal by deep learning greatly improves the accuracy rate of diagnosing the pathological voice while comparing with other state of art.

REFERENCES

[1] K. Verdolini and L. O. Ramig, "Occupational risks for voice problems," Logopedics Phoniatrics Vocology, vol. 26, no. 1, pp. 37-46, 2001.

[2] [2] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society, Engineering in Medicine and Biology, 2002, vol. 1, pp. 182-183 vol.1.

[3] M. K. Arjmandi and M. Pooyan, "An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine," Biomedical Signal Processing and Control, vol. 7, no1, pp. 3-19, 2012/01/01/ 2012.

[4] M. Hariharan, K. Polat, and S. Yaacob, "A new feature constituting approach to detection of vocal fold pathology,"International Journal of Systems Science, vol. 45, no. 8, pp. 1622-1634, 2014/08/03 2014

[5] A. Al-nasheri et al., "An Investigation of Multidimensional VoiceProgram Parameters in Three Different Databases for VoicePathology Detection and Classification," Journal of Voice, vol.31, no. 1, pp. 113.e9-113.e18.

[6] G. Muhammad et al., "Voice pathology detection using interlacedderivative pattern on glottal source excitation,"

Biomedical SignalProcessing and Control, vol. 31, pp. 156-164, 2017/01/01/ 2017.
[7] G. E. Hinton, S. Osindero, Y. Teh, "A fast learning algorithm for deep belief nets.," Neural Comput., vol. 18, pp. 1527–1554, 2006.
[8] A. Kae, K. Sohn, H. Lee, E.Learned-Miller, "Augmenting CRFs with Boltzmann Machine Shape Priors for Image Labeling.," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[9] Maria Markaki, Yannis Stylianou,Voice Pathology Detection and Discrimination Based on Modulation SpectralFeatures, IEEE Transactions on Audio Speech and Language Processing, pp 1-12, October 2011
[10] N. Malyska, T. Quatieri, and D. Sturim, "Automatic dysphonia recognition using biologically inspired amplitude-modulation features," in Proc.ICASSP, 2005, pp. 873–876.
[11] Markaki, M., Stylianou, Y. Voice Pathology Detection and Discrimination Based on Modulation Spectral Features. IEEETransactions on Audio, Speech, and Language Processing, Vol. 19, No. 7, 1938-1948, 2011.
[12] Panek, D., Skalski, A., Gajda, J., Tadeusiewicz, R. Acoustic Analysis Assessment in Speech Pathology Detection. Int. J. Appl. Math. Comput. Sci., 2015, Vol. 25, No. 3, 631–643.
[13] Al-nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z. Investigation of Voice Pathology Detection and Classification on Different Frequency Regions Using Correlation Functions. Journal of Voice, 31(1):3-15, 2016
[14] Cordeiro, Hugo T. Reconhecimento de Patologias da VozusandoTécnicas de Processamento da Fala. PhD thesis at Universidade Nova de Lisboa, 2016.
[15] Barry, W.J., Pützer, M. Saarbrücken Voice Database, Institute of Phonetics, Univ. of Saarland, http://www.stimmdatenbank.coli.unisaarland.de/