www.revistageintec.net
ISSN: 2237-0722
revista GEINTEC Gestão, Inovação e Tecnologias

Enhanced Optimization in DCNN for Conversion of Non Audible Murmur to Normal Speech based on Dirichlet Process Mixture Feature

T. Rajesh Kumar[1]; Arumbaka Srinivasa Rao[2]; K. Kalaiselvi[3]; S.S. Manivannan[4]; C. Shahul Hameed[5]; C.M. Velu[6]

[1]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India.

[1]t.rajehs61074@gmail.com

[2]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India.

[2]sri.srinivas.07@gmail.com

[3]Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India.

[3]mkkalai1981@gmail.com

[4]School of Information Technology & Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.

[4]manivannan.ss@vit.ac.in

[5]Department of CSE, Saveetha School of Engineering, Saveetha University, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India.

[5]shahulfriend@gmail.com

[6]Department of CSE, Saveetha School of Engineering, Saveetha University, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India.

[6]velucm.sse@saveetha.com

Abstract

In various applications towards mobile communication, the recognition of speech automatically plays a vital role. The user communication devices for interaction require a large amount of vocabulary recognition system, more preciseness and real time less power consuming schemas. The miniaturized battery controlled devices suffer with power consumption and large memory bandwidth. Also the speech challenge people's mobile devices require more attention. Therefore, a useful technology is proposed to convert non-audible murmur to normal speech, based on Stochastic Biogeography-based WOA (SBWOA) integrated with Dirichlet process mixture. The features like spectral skewness, Taylor AMS, spectral centroid, pitch chroma and newly developed Dirichlet Process mixture features

Received: 25.05.2021 – Accepted: 23.06.2021

are extracted from the input murmured speech signal and trained in the DCNN. The identification of speech is based on Deep Convolutional Neural Network (Deep CNN), which is trained by the proposed Stochastic Biogeography WOA (SBWOA). The stochastic gradient descent method, Biogeography-based optimization (BBO) and Whale optimization algorithm (WOA) are combined together to improve the results in metric analysis. The TRP, FPR and Accuracy shows the improved results of 0.9, 0.001 and 0.99 respectively.

Index Terms: Biogeography-based Optimization, Deep Convolutional Neural Network, Dirichlet Process Mixture, Speech Recognition, Stochastic Gradient Descent Approach.

1. Introduction

For the translation of audio signals or human speech to text, ASR systems are used. Furthermore, in many implementations, such as spoken text extraction, Integrated Voice Response (IVR) programmes, and dictation interfaces, ASR is used. For the above-mentioned applications, however, a clear speech recognition system is needed. Moreover, notable advances have been made in some of the frameworks for automatic speech recognition[1][35]. The ASR systems have been shown optimum efficiency in the modern day, while on the other hand; there are many limitations, challenges and shortcomings that are solved by improved and successful human-computer interaction quality[33][26]. Despite much of the research that has plagued the growth of the ASR systems over the last few decades, but even under certain circumstances, both of these systems remain largely unstable and sensitive. In addition, ASR performance is determined by many influences, including the nature of input speech and forms and actual speaker features, such as accent, style and speech intensity, psycho-physical state of the speaker, anatomy of the vocal tract, etc.[6][32].

In speech technology, Whisper plays an important role because of its substantial variance relative to the normal phoned speech, which induces noisy form, absence of glottal sounds, and less Signal to Noise Ratio (SNR)[6][29]. Murmur is a basic verbal communication system that is widely used under different situations. Initially, it is used in the discourse to establish a personal and discreet atmosphere and then to shield a variety of private and sensitive details from uncomplicated parties[28][36]. Whispered speech is created by adjusting the vocal cords while engaged in glottis development of constricted restraint, inducing vocal area excitation[31]. It is necessary to learn the procedure of embedding data in speech without the existence of critical frequency in speech invention[9][8]. In different cases, the ASR system runs, but it is still not possible to monitor the speech method of the speaker in real world use. It thus induces inequalities between the conditions of research and training[6][24].

Due to recent advances in computer hardware technologies and machine learning methods, the Deep Neural Network (DNN) is used successfully in the ASR framework. In addition, Secret Markov Model (HMM) [25] with DNN approaches (DNN-HMM) was used in various studies to build series of audio observations augmented by HMM[37] with Gaussian Mixture Model (GMM-HMM) in different occupations, such as massive vocabulary and very large scale datasets[7][37]. In addition, different techniques have been developed to boost audio divergence through model adaptation[10-14][8], other sensing approaches, such as throat microphone[15][8][27], and transformation of features[11][8][30]. Audio-visual technique was also applied here to isolate word recognition under the condition of murmur speech[16][8]. The traditional features of the Mel-Frequency Cepstral Coefficients (MFCC), GMM-HMM method or Perceptual Linear Prediction (PLP)[8] have been established. Moreover, for the recognition of speech[35][34], unsupervised learning of spare spectro-temporal sound coding was applied. Along with this, learning of usual codes or sounds in acoustic cortex was applied in unsupervised method. The supervised data driven technique utilized Linear Discriminant Analysis (LDA) to obtaining temporal modulation filters [3].

The purpose of the thesis is to establish a system for converting a non-audible whisper into ordinary speech. The steps followed in the developed model are pre-processing, feature extraction, and speech recognition. Based on scanning, the input speech signal is initially pre-processed. Once the signal is pre-processed, the extraction of features is carried out to remove suitable features for further processing. The speech recognition is done using Deep CNN, which is educated by SBWOA, after feature extraction.

The main contributions of the paper are:

- **Proposed SBWOA-based Deep CNN for speech recognition**: To train Deep CNN for speech recognition, the proposed SBWOA algorithm is employed. The SBWOA is a variation of the solution to stochastic gradient descent, WOA, and BBO. In order to improve device performance, the developed SBWOA-based Deep CNN is optimised for speech recognition. Moreover, for collecting the characteristics for further processing, the Dirichlet process mixture is added.

The residual sections of paper are pre-arranged as below: Section 2 illustrates depiction of conventional speech recognition methods developed in literature and faced problems, which are included as literature review to design proposed method. The developed approach for speech recognition based on Dirichlet process mixture and SBWOA-based Deep CNN is explained in Section 3. The outcomes of developed strategy with other approaches are displayed in Section 4 and finally, Section 5 presents conclusion.

2. Literature Review

Some of the existing speech recognition methods with their drawbacks are mentioned in this chapter, which inspire researchers to create a new approach for speech recognition. The ten classical techniques focused on understanding of speech and its limits are discussed below:

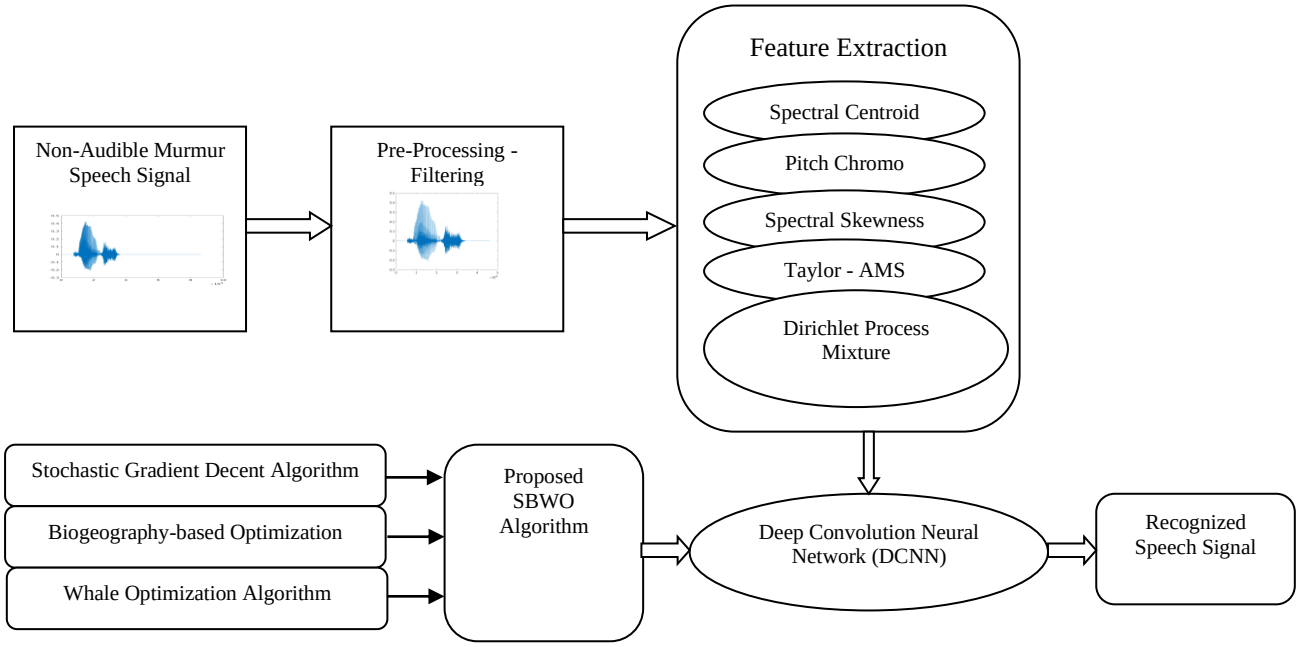
An approach to mitigating error rate detection has been developed by Toktam Zoughiet al. [1]. Initially, the Adaptive Windows Convolution Neural Network (AWCNN) was applied here to study the variations in common temporal and spectral characteristics. In addition, residual learning has been modified to provide increased control over the relocation of input results. In order to provide optimum information flow to deeper levels, the approach failed to consider other applications, such as Text to Speech (TTS), image, ECG and EEG, signal processing. Rongfeng Suet al.[2] introduced the Audio Visual Speech Recognition (AVSR) system's cross-domain deep visual feature development tool. In order to achieve acoustic-to-visual (A2V) inversion techniques, Bidirectional Long Short-Term Memory (BLSTM) and 3D Audio-Visual Mandarin Continuous Expression (3DAV-MCS) were implemented on the basis of the individual statement. Audio domain divergence between the 3DAV-MCS AV aim and parallel data in which the A2V inversion was mitigated by multi-level adaptive deep neural network by cross-domain modification process multi-level adaptive deep neural networks. In addition, A2V inversion and cross-domain adaptation integration have enabled the best visual characteristics of acoustic data from the mismatch domain. However, for the purpose of deciding the visual function of the Evert time phase, long-term acoustic knowledge was not considered. Purvi Agrawa let al.[3] generated a method focused on deep variation models for modulation filter learning. In addition, to overcome the problems of filter learning, and to catch large voice modulations, the deep unsupervised generative modeling method was added. The solution requires to train the two separate ASR systems for ASR enhancements on each function stream. An approach to reducing energy consumption and memory needs was developed by Reza Yazdani et al.[4]. Initially, the Locality-AWare-Scheme (LAWS) was implemented between consecutive segments to exploit locality. The reliability of the LAWs was then increased by adapting the workload of ASRs using run-time feedback. In order to maintain the ASR workload, the method neglected to recognize smaller classes to have greater granularity. Thomas Hueber and Gérard Baillya[5] developed HMM full-covariance for the show of HMM-enabled consonant segmental intelligibility. This form, however, does not involve articulatory differences between silent and modal voice.

Grozdić et al. [6] used deep de-noising auto-encoder to substantially minimize inter-speaker standard deviation of word recognition accuracy in the unequal case, but features struggled to display any relevance that materialized as a drawback of this scheme in neutral circumstances. Shabnam Ghaffarzadegan et al. [7] modelled a deep neural network to classify speech in order to reduce the phoneme error rate. In this case, the approach considered pseudo-samples for the training and allowed a short training time, but the training technique is not compactable because of less training details. Slobodan T. Jovicic and Đor'e T. Grozdic[8] used Deep De-noising Auto-encoder and inverse filtering to eliminate whisper and neutral voice variance with increased rates of word recognition in incompatible contexts. Wentao Fan et. al.[9] designed algorithms for extraction of features from speech signal by applying Dirichlet process mixture and used a batch variational learning of the model. Francois Caron and others[10] discussed about the DPM for Bayesian Inference for Dynamic models for training the algorithms such as Monte Carlo Markov Chain, Rao-Blackwellisation for designing Kalman filters based on particle filter. The limitation of this approach is that in the murmur or neutral circumstance, when ASR system anticipates both the unvoiced and the voiced sounds.

3. Proposed Dirichlet Process Mixture based Linear Predictive for Speech Recognition

Speech Recognition transforms the way people communicate on handheld devices –mobiles with diverse apps. However, satisfying user interactivity requires not only a very broad, detailed framework for vocabulary detection, but also an energy-efficient, real-time structure. In order to produce non-audible whisper to ordinary voice, an efficient procedure, called Dirichlet process mixture, is therefore created. Initially, the input speech signal is pre-processed by filtering, and the required characteristics are extracted in the function extraction procedure. The system built by the introduction of the Taylor series into the regular AMS is Taylor-AMS. Deep CNN[21], in which the training algorithm is performed using Stochastic Biogeography-based WOA Optimization, which is built using stochastic gradient descent algorithm[22], WOA[23], and BBO[22], performs speech recognition after the feature extraction method. The developed model then classifies the speech signal into ordinary speech based on extracted characteristics. The figure 1 depicts the schematic diagram of proposed SBWOA-DPM for speech recognition.

Figure 1 - Schematic diagram of proposed SBWOA-DPM for speech recognition



A. The Input Signal is Received

By considering the signals gathered from the dataset, the speech signal is known. Let us assume the dataset as Y defined as the n number of input speech signals and as,

$$\mathrm{Y} = \{C_i\}\,;\; i \in \{1,2,...,n\} \qquad (1)$$

where, Y indicates dataset, C specifies the input speech signal, and n signifies total amount of speech signals present in dataset. However, the input video C_i is selected from the dataset to perform the speech recognition.

B. Using filtering, Pre-processing been done

To allow smooth processing of an input speech signal, the input speech signal C_i is selected and given to the pre-processing level. Here, the Wiener filter and elimination of silence are used to pre-process speech signal data. The silence deduction, however, is the pre-processing procedure for eliminating silence and the voiceless divisions of the speech signal and the Wiener filter is used to filter noise from the input signal and C_i^* is denoted as the pre-processed output.

C. After Pre-processed signal, feature extraction done

The pre-processed signal is C_i^* forwarded to the feature extraction module until the speech signal is pre-processed. The feature extraction is done here on the basis of Taylor-AMS, pitch chroma, spectral centroid, spectral skewness, and the mixture of the Dirichlet method to produce highly important features for improved speech signal recognition. In addition, the signal complexity is minimised, thus reducing the range of characteristics. Furthermore, the precision associated with classification ensures efficient extraction of the feature. As below, the function extraction phase followed in this paper is outlined. Furthermore, the precision associated with classification ensures efficient extraction of the feature. As below, the function extraction phase followed in this paper is outlined.

a) Spectral Skewness

In order to calculate the distribution symmetry, it applies to the coefficient spectrum skewness. The skewness coefficient is measured as the ratio of skewness to the third power of standard deviation and the coefficient of skewness is expressed by,

$$Skewness = \frac{\sum_{\mu=1}^{\kappa}\left[(\kappa - \nu)^3 \times \varpi\right]}{\alpha^3} \qquad (2)$$

where, the term κ refer to individual spectrum, ν signifies spectrum mean of input signal, ϖ signifies spectrum width of input signal, and the term α indicates standard deviation. The output of spectral skewness is denoted as u_1.

b) Pitch Chroma: The pitch chroma of input speech signal is computed using the below expression,

$$B_a(\kappa) = \sum_{j=0}^{d-1} \left| R_{ba}(\kappa + j\beta) \right| \qquad (3)$$

where, R_{ba} signifies log-frequency spectrum, the term j refer to numeral octave index where $j \in [0, d-1]$. The octaves are indicated as, d and κ represents chroma index or integer pitch class, denoted as $\kappa \in [0, \beta - 1]$ such that β represents bins per octave. Here, octave is defined as the

distance of 12 pitches in the individual separated pitch with the two elements, which includes chroma tone and height. Therefore, the chroma feature is aggregated the information in simple manner that relates pitch class as single coefficient. Hence, the pitch chroma output is indicated as u_2.

c) Spectral centroid: Itrefertocenter frequency where spectral energy centers is elevated, and is prepared using spectral level of frequency bin in an frequency and input signal at certain bin. Therefore, spectral centroid is expressed by,

$$SC(\mu) = \frac{\sum_{\mu=1}^{\rho} \mu . \Re(\mu)}{\sum_{\mu=1}^{\rho} \Re(\mu)} \qquad (4)$$

Where, the total bins are denoted as ρ. Here, the spectral centroid represents center of gravity of input speech signal, and is normalized based on minimal centre of frequency $\Re(\mu)$. The output spectrum centroid is indicated as u_3

d) Taylor-AMS Features:

AMS[21, 26] attributes are paired with the Taylor series to define historical classification criteria for extracting Taylor-AMS characteristics. Here, except in the case of noisy signals, the AMS function provides robust speech recognition with useful knowledge. Furthermore, AMS features include the phase and log-magnitude information of the input signal. Furthermore, AMS features include the phase and log-magnitude information of the input signal. The combination of the Taylor series with extracted AMS features guarantees improved classification in order to retrieve features using past data features. In Taylor-AMS, the processing steps taken are frame manufacturing, windowing, FFT, and triangular feature production.

i) Read input signal: Input speech signal is given to easy pre-processing steps, such as quantization and sampling before feature extraction stage, the, hence signal is made fit for further processing.

ii) Application of band pass filter for generating time-frequency channels: At first, band pass filters are executed to dissolve noisy signal as the time-frequency divisions resulting 85 framing divisions and every section represents the channel.

The band pass filter activates pre-arranged frequency input signals and terminates other signals reflecting individual paths carrying lower and upper bound frequencies.

iii) Rectification-Determination of envelop:

The rectification aims to arrange the encircle for the destroyed character channels on the basis of aspect three, which for the 128 samples forms 64 overlapping parts. Then, determined parts are given to the window centred on the Hanning window that decreases unnecessary signals, enabling the input signal to gather realistic information.

iv) Windowing concept for Signal framing: The infinite input signal flow is converted into constant streams, which are called as frames in order to maximize the speech signal, and the framing phase allows stationary signal products to prevail. In the framing portion, the edge has the potential to add signal harmonics. Therefore, in the contrasting modes of the frames that differ between themselves, the fine tuning continues to make improvements. Then, the second frame distributes half of the previous frame and half of the subsequent frame, protecting the edge level detail. The Hann window is used for windows, and $[1\times 255]$the size of the window is0.5, along with the overlap rate.

v) Extraction of Taylor-AMS Feature: In order to find the spectrum modulation of frames, FFT is given the framing output. For generating AMS functions, the FFT output and the triangular form windows are replicated. Assume the vector of AMS features as with dimension. In addition, the minor alternations in the frequency and time domain of both signals are calculated to form delta AMS functions.

Assume the vector of AMS features as $G(s,a)$with dimension $[255\times 85]$. In addition, the minor alternations in the frequency and time domain of both signals are calculated to form delta AMS functions. In addition, the minor alternations in the frequency and time domain of both signals are calculated to form delta AMS functions as,

$$G(s,a)=\left[G(s,a),G_Q(s,a),\Delta D_k(s,a)\right] \quad (5)$$

$$\Delta G_Q(1,a)=G(2,a)-G(1,a) \text{ when } s=1 \quad (6)$$

$$\Delta G_Q(s,a)=G(s,a)-G(s-1,a) \text{ when } s=2,...,Q \quad (7)$$

where, the term $\Delta G_Q(s,a)$signifies delta feature vector, the term Q represents total segments, $G(s,a)$represents AMS feature, and Delta-AMS feature at the time 1 sec is illustrated in equation (5). In general, delta AMS features are expressed in equation (7). The final equation of Taylor-AMS feature is given by,

$$G(s,a)=\Delta G_Q(s,a)+\left[\begin{array}{l}2G(s)-2.7182G(s-2)+2.718G(s-3)-1.359G(s-4)+0.4518G(s-5)-\\0.111G(s-6)+0.0208G(s-7)-0.00276e^{-3}G(s-8)+0.00019G(s-9)\end{array}\right] \quad (8)$$

Taylor-AMS function output is defined as, u_4.

Received: 25.05.2021 – Accepted: 23.06.2021

e)Proposed Dirichlet Process Mixture

In particular, the hierarchical Dirichlet process structure is useful for clustered data modeling problems where observations are divided into groups that allow the sharing of mixture components to remain statistically related[10]. A hierarchical Dirichlet process mixture model of generalized Dirichlet distributions with an unsupervised feature selection scheme is built. For the input non audible murmur signal, when the probability distribution over the signal be 'H' and the random distribution be 'G' in the space of probability 'θ' with a real positive number be 'α', then

$$G \approx DP(\alpha, H) \tag{9}$$

If a_1, a_2, …a_t are the set of finite samples in the free space of 'θ' in the murmured signal, then

$$G(a_1), G(a_2), ..G(a_t) \approx DP(\alpha H(a_1), (\alpha H(a_2), ...(\alpha H(a_t)) \tag{10}$$

Whereas right hand side of the above equation(10) is a finite-dimensional Dirichlet process distribution with all parameters. By applying the hierarchical Dirichlet process based on Bayesian model [9] for creating the model based clustering of the input data set, also using stick-breaking representation for nonparametric coefficients are as follows:

$$\theta_{t+1} \mid \theta_t \approx \frac{1}{t+\alpha} \sum_{k=1}^{t} \delta_{\theta_k} + \frac{\alpha}{t+\alpha} H_0 \tag{11}$$

Whereas δ_{θ_k} be the delta function. The previous distribution of π is a particular Beta distribution[9], then

$$p(\pi) = \prod_{j=1}^{M} \prod_{t=1}^{\infty} Beta(1, \lambda_{jt}) = \prod_{j=1}^{M} \prod_{t=1}^{\infty} \lambda_{jt} (1-\pi_{jt})^{\lambda_{jt}-1} \tag{12}$$

Because of its high performance in modelling murmured speech signal's high dimensional data set, using generalized dirichlet (GD) distribution with feature selection is observed as Hierarchical Dirichlet process mixture model distributions. In addition, a general covariance structured is implemented in dirichlet process distributions.

The GD distribution having the arguments of $\vec{\alpha} = \alpha_1, \alpha_2 .. \alpha_d$ and $\vec{\beta} = \beta_1, _2 . \beta .. \beta_d$ with d-dimension random vector of $\vec{X} = X_1, X_2 ... X_d$ then

$$GD(\vec{X} \mid \vec{\alpha}, \vec{\beta}) = \prod_{k=1}^{d} \frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} X_k^{\alpha_1 - 1} (1 - \sum_{f=1}^{k} X_f)^{\gamma_1} \tag{13}$$

Whereas

$\sum_{k=1}^{d} X_k < 1; 0 < X_k < 1 for k = 1,2,..d; \alpha_k > 0, \beta_k > 0, \gamma_k = \beta_k - \alpha_{k+1} - \beta_{k+1}; for all k = 1,2,..d-1, \gamma_d = \beta_d - 1.$ Also $\Gamma(x)$ is a gamma function. By implementing mathematical property of general dirichlet distribution, the alpha function, Beta functions and its corresponding likelihood function, with prior probability

Received: 25.05.2021 – Accepted: 23.06.2021

distribution of $\vec{\phi}$ may be defined with the relevant features of X_{jil} and Bernoulli's variable of $\vec{\phi}_{jil}$ are as follows:

$$p(\vec{\phi} \mid \vec{\varepsilon}) = \prod_{j=1}^{M} \prod_{i=1}^{N} \prod_{l=1}^{d} \varepsilon_{l_1}^{\phi_{jil}} \varepsilon_{l_2}^{1-\phi_{jil}} \quad (14)$$

Whereas the vectors $\vec{\varepsilon} = \vec{\varepsilon}_1, \vec{\varepsilon}_2, .. \vec{\varepsilon}_d$ represents the features, such that $\vec{\varepsilon}_{l_1} + \vec{\varepsilon}_{l_2} = 1; \vec{\varepsilon} = \vec{\varepsilon}_{l_1}, \vec{\varepsilon}_{l_2}$ is a probability dirichlet process mixture, also ϕ_{jil} is one of the Bernoulli variables with $p(\varphi_{jil} = 1) = \varepsilon_{l_1}; p(\varphi_{jil} = 0) = \varepsilon_{l_2}$. In addition the dirichlet process mixture over $\vec{\varepsilon}$ will be given as following equation:

$$p(\vec{\varepsilon}) = \prod_{k=1}^{d} Dir(\vec{\varepsilon}_k \mid \vec{\xi}) = \prod_{k=1}^{d} \frac{\Gamma(\xi_1 + \xi_2)}{\Gamma(\xi_1)\Gamma(\xi_2)} \varepsilon_{k_1}^{\xi_1 - 1} \varepsilon_{k_2}^{\xi_2 - 1} \quad (15)$$

The likelihood function also defined with the same ξ vectors equation, which leads to have the features of input speech signal based on dirichlet process mixture. The extracted feature of dirichlet process mixture is denoted as u_5.

For speech recognition, the characteristics are derived from the feature extraction stage and are then integrated into the feature vector. The functionality factor is represented as,

$$U^{final} = [u_1, u_2, u_3, u_4, u_5] \quad (16)$$

where, the extracted features are represented as u_1, u_2, u_3, u_4 and u_5. The size of each technical indicator is $[1 \times 1]$, and the extracted feature is of dimension $[1 \times 6]$.

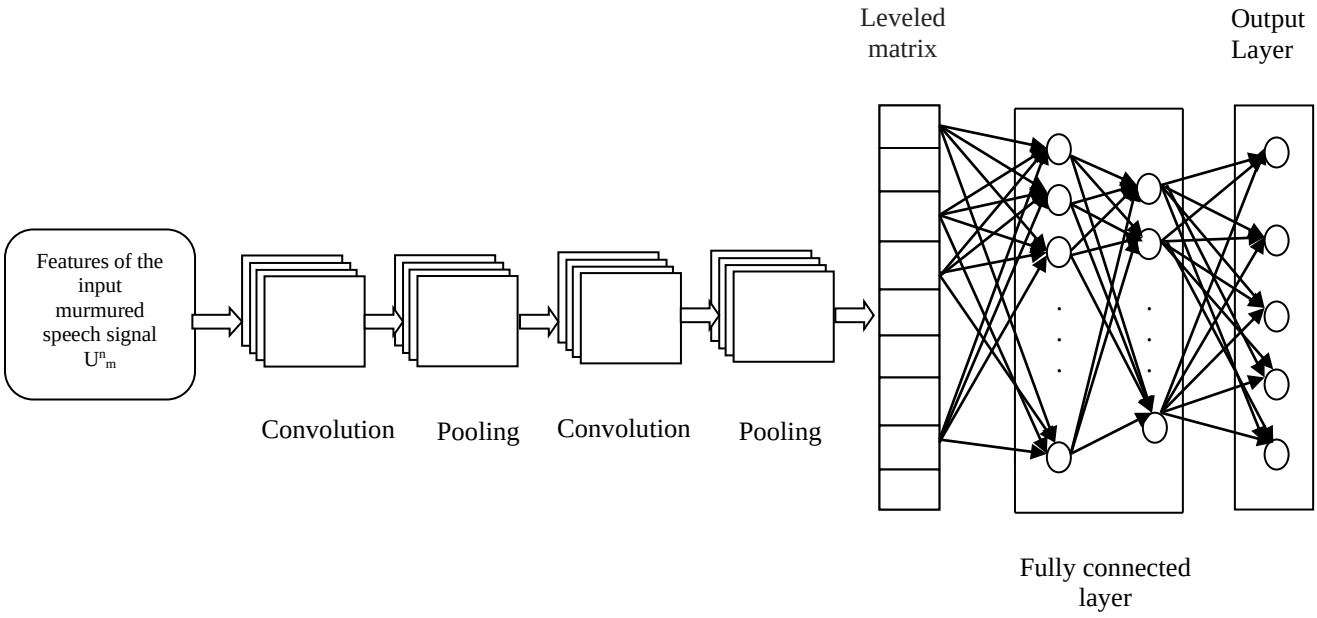
D. Alteration of non-audible murmur to normal speech based one Proposed SBWOA-based Deep CNN

Once the appropriate features get extracted, speech recognition is done based on Deep CNN [21]. The feature vector is taken as input of Deep CNN for the conversion of non-audible murmur to a regular speech. However, the Deep CNN tunes the biases and weights of classifier for improving the classification accuracy. In addition, the Deep CNN is trained by developed optimization method, named SBWOA that is designed newly by incorporating stochastic gradient descent algorithm, WOA, and BBO.

a) Deep CNN Architecture

Deep CNN [21,26]is utilised to convert the non-audible murmur to the normal speech to achieve optimal speech recognition output. The Deep CNN architecture consists of three separate layers, such as the convolutionary layer, the pooling layer, and the completely linked layer. Each layer executes its own operations in the Deep CNN architecture. Each layer executes its own operations in the Deep CNN architecture. The convolutionary layer is used here to measure the function diagram, and then the pooling layer performs sub-sampling. Finally, using the completely linked layer, the classification is performed. Classification precision is strengthened as the convolutional layers in the Deep CNN classifier are expanded. In addition, the first-layer neuron is paired with the following layer of individual neurons. The form of the Deep CNN classifier is seen in Figure 2.

Figure 2 - Architecture of Deep CNN



Convolutional layer: Through considering the covolutionary filter that links neurons from the previous layer to the next layer by trainable weights, the convolutional layer is used to generate patterns for femur points. Assume that the convolutional layer input is expressed as the convolution layer output,

$$\left(U_m^n\right)_{g,h} = \left(T_m^n\right)_{g,h} + \sum_{z=1}^{e_1^{z-1}} \sum_{v=-e_1^j}^{e_1^n} \sum_{x=-e_2^j}^{e_2^n} \left(\varpi_{m,z}^n\right)_{x,y} * \left(F^{\omega}\right)_{g+v,h+x} \qquad (17)$$

where, $\left(U_m^n\right)_{g,h}$ refer to the output of n^{th} convolutional layer or the feature map, and $*$ refer to convolutional operator. $\varpi_{m,z}^n$ signifies the weight, and the term T_m^n indicates the bias of n^{th} convolutional layer.

ReLU layer: For applying the non-saturating activation function, ReLU is the rectified linear unit. Here, by setting them to zero, the negative values are efficiently extracted from the activation diagram. As the input of the next k^{th} convolutionary layer, the output generated by $n-1^{th}$ layer is forwarded and expressed as,

$$U_m^n = A\left(U_m^{n-1}\right) \qquad (18)$$

where, the activation function is indicated as A.

Pooling layer: The pooling layer is often considered a non-parametric layer, and is used to execute a fixed operation that ignores bias and weight consideration.

Fully connected layers: The pooling layer output is taken as the input of the fully linked layer and the output from the fully linked layer is given as,

$$V_m^n = \sigma\left(U_m^{n-1}\right) \text{ with } U_m^n = \sum_{z=1}^{e_1^{z-1}} \sum_{v=-e_1^j}^{e_1^n} \sum_{x=-e_2^j}^{e_2^n} \left(\varpi_{m,z}^n\right)_{x,y} * \left(F^{\omega}\right)_{g+v,h+x} \qquad (19)$$

where, the term $\left(\varpi_{m,z}^n\right)_{x,y}$ refer to weights. The weight values are computed for tuning Deep CNN such that the optimal weight is retrieved.

b) Training of Deep CNN based on Stochastic Biogeography-based WOA

The preparation of the Deep CNN classifier is performed to achieve a better weight factor using the established SBWOA. The SBWOA is a mixture of stochastic gradient descent techniques[22], WOA[23], BBO[24] for successful collection of optimum weights. The parametric characteristics obtained from the optimization algorithm described above allow efficient classification efficiency based on the input image. The stochastic algorithm is efficient since it is linear for training data and is capable of approximating the true gradient over time for each training data. WOA, on the other hand, is stimulated by the hunting strategy of humpback whales in which the search for its prey is based on the mechanism of bubble-net attack. In addition, by sequence of stages, such as exploitation, discovery, and encircling, the search for prey is related in or to the outer search spaces. However, the faced by the above cites two algorithms are solved by the BBO algorithm.

Biogeography is an analysis related to the distribution of biological species through geography. In the 1960s, the mathematical equations that manage the allocation of species were initially developed and discovered. Furthermore, biography promotes solving many problems in optimization. Thus, the stochastic WOA update equation is updated using BBO. This move then helps the approach, with better efficiency, to be more effective. Below is the algorithmic method of the proposed SBWOA.

Step 1: Initialization: The zero vectors are initialized in the first step, which is represented as Z, and select the training sample randomly represented as, $\left(H_l^h, M_l^h\right)$. The term H_l^h denotes the feature vector of l^{th} training sample with dimension η, and the term M_l^h refers to category of training data.

Step 2: Fitness function computation: The fitness is estimated for finding optimal solution for speech recognition. Moreover, it is estimated by minimum error value, and solution with respect to minimal error is taken as best solution. The fitness is formulated using the below expression,

$$MSE = \frac{1}{\kappa}\left[\sum_{o=1}^{\kappa} U_{target} - U_m^n\right] \qquad (20)$$

where, κ refer to training samples, the terms U_{target} and U_m^n signifies the target and estimated output of classifier.

Step 3: Update the solution using proposed SBWOA:

The weights are calculated using the suggested SBWOA for training Deep CNN after the fitness measurement, and updating is performed using weights that donate to minimum error value. The stochastic WOA's weight update equation is expressed by,

$$Z_{h+1} = \begin{cases} Z_{h+1}^{SGD} & ; \quad if \ c_{avg}^{SGD} < c_{avg}^{WOA} \\ Z_{h+1}^{WOA} & ; \quad Otherwise \end{cases} \qquad (21)$$

$$Z_{h+1}^{SGD} = \left(1 - \frac{1}{h}\right) Z_h + M_l^h \times H_l^h \qquad (22)$$

where, the term h refers to iteration, Z_h signifies optimal solution in previous iteration, and Z_{h+1} refer to optimal solution in next iteration. The standard equation of BBO is given by,

$$Z_{h+\Delta h}^e = Z_h^e\left(1 - \gamma_e \Delta h - \nu_e \Delta h\right) + Z^{e+1}\gamma_{e+1}\Delta h + Z^{e+1}\nu_{e+1}\Delta h \qquad (23)$$

$$Z_h^e = \frac{Z_{h+\Delta h}^e - Z^{e+1}\gamma_{e+1}\Delta h - Z^{e+1}\nu_{e+1}\Delta h}{1 - \gamma_e \Delta h - \nu_e \Delta h} \qquad (24)$$

Received: 25.05.2021 – Accepted: 23.06.2021

Substituting in (22),

$$Z_{h+1}^{SGD} = \left(1 - \frac{1}{h}\right)\left(\frac{Z_{h+\Delta h}^{e} - Z^{e+1}\gamma_{e+1}\Delta h - Z^{e+1}\nu_{e+1}\Delta h}{1 - \gamma_e \Delta h - \nu_e \Delta h}\right) + M_l^h \times H_l^h \tag{25}$$

The above equation is the updated equation during SGD based on BBO. Then, the standard equation of WOA is expressed by,

$$Z_{h+1}^{WOA} = \begin{cases} Z_h^* - \vec{U}.\vec{N} & ; \quad if\ r < 0.5 \\ \vec{\kappa}^{'} e^{pq}.\cos(2\pi q) + \vec{Z}_h^* & ; \quad if\ r \geq 0.5 \end{cases} \tag{26}$$

where, $\vec{\kappa} = \left|\vec{Z}_h^* - \vec{Z}_h\right|$ Substituting $\vec{Z}_h$ from BBO,

$$\vec{\kappa} = \left|\vec{Z}_h^* - \frac{Z_{h+\Delta h}^{e} - Z^{e+1}\gamma_{e+1}\Delta h - Z^{e+1}\nu_{e+1}\Delta h}{1 - \gamma_e \Delta h - \nu_e \Delta h}\right| \tag{27}$$

$$Z_{h+1}^{WOA} = \begin{cases} \vec{Z}_h^* - \vec{U}.\vec{N} & ;if\ r < 0.5 \\ \left|\vec{Z}_h^* - \frac{Z_{h+\Delta h}^{e} - Z^{e+1}\gamma_{e+1}\Delta h - Z^{e+1}\nu_{e+1}\Delta h}{1 - \gamma_e \Delta h - \nu_e \Delta h}\right| e^{pq}.\cos(2\pi q) + \vec{Z}_h^* & ;if\ r \geq 0.5 \end{cases} \tag{28}$$

where, the term Z_h^* signifies optimal location of whale in previous iteration, $\vec{U}$, and $\vec{N}$ signifies coefficient vectors, which are using constants, λ_1 and λ_2, respectively. The constant λ_1 linearly reducing from 2 to 0 that indicates switching among exploitation and exploration phases, and constant λ_2 indicates random vector that ranges among 0 and 1. The term p denotes constant, which describes logarithmic spiral form and q constitutes random number in 0 and 1. The term $\vec{\kappa}^{'}$ represents the distance among whale and prey.

Step 4: Recheck the feasibility: Based on fitness value, the feasibility is again computed, hence if the generated new solution is best than previous one, the old one is replaced by the new solution.

Step 5: Termination: The steps described above are repetitive before the best solution for speech recognition is obtained. Therefore, the speech signal characteristics are collected when the new signal arrives for speech recognition, and are forwarded to the classifier to process the characteristics in terms of new function dataset and create class mark for input speech signal. In recognizing words that are subject to many uses, the evolved speech recognition technique is very successful.

4. Results and Discussion

This section addressed the findings of the evolved Dirichlet Process Mixture-based SBWOA algorithm for producing natural speech from non-audible murmurs.

a) Experimental setup

The experimentation of developed model is developed in MATLAB tool with the windows 10 OS, 4GB Ram, and the intel I3 processor.

b) Description of database

TIMIT[25] is used to conduct the experiment, and the overview of the datasets is given below: The TIMIT corpus[25] includes a read speech to enable the automated processing of the speech recognition system. TIMIT consisted of 630 speakers of speech signal, which is rich in phonetics, orthographic time-aligned, and 16-bit speech waveform word transcriptions of 16kHz for the actual utterance. The TIMIT amount transcripts are checked with the preparation and evaluation sub-sets with the required dialectal coverage and phonetics criteria.

c) Evaluation metrics

Using four metrics, such as accuracy, True Positive Rate (TPR), and False Positive Rate (FPA), Receiver Operating Characteristic (signal detection theory- ROC), the efficiency of the established approach is evaluated.

a) Accuracy: It is used to measure the degree of closeness of the approximated value to its real value in the optimal classification of big data, which is expressed by,

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{EP} + \mathrm{EN}} \qquad (29)$$

where, true positive is denoted *as TP, EP refer* to false positive, *TN* signifies the true negative and the false negative is indicated as *EN*, respectively.

b) TPR: TPR measures true positives that are precisely distinguished by the method developed. The sensitivity word is,

$$TPR = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{EN}} \qquad (30)$$

where, TP signifies true positive and FN is the false negative.

c)FPR: It displays real negatives correctly recognized on the basis of an existing classifier. The specificity is conveyed as,

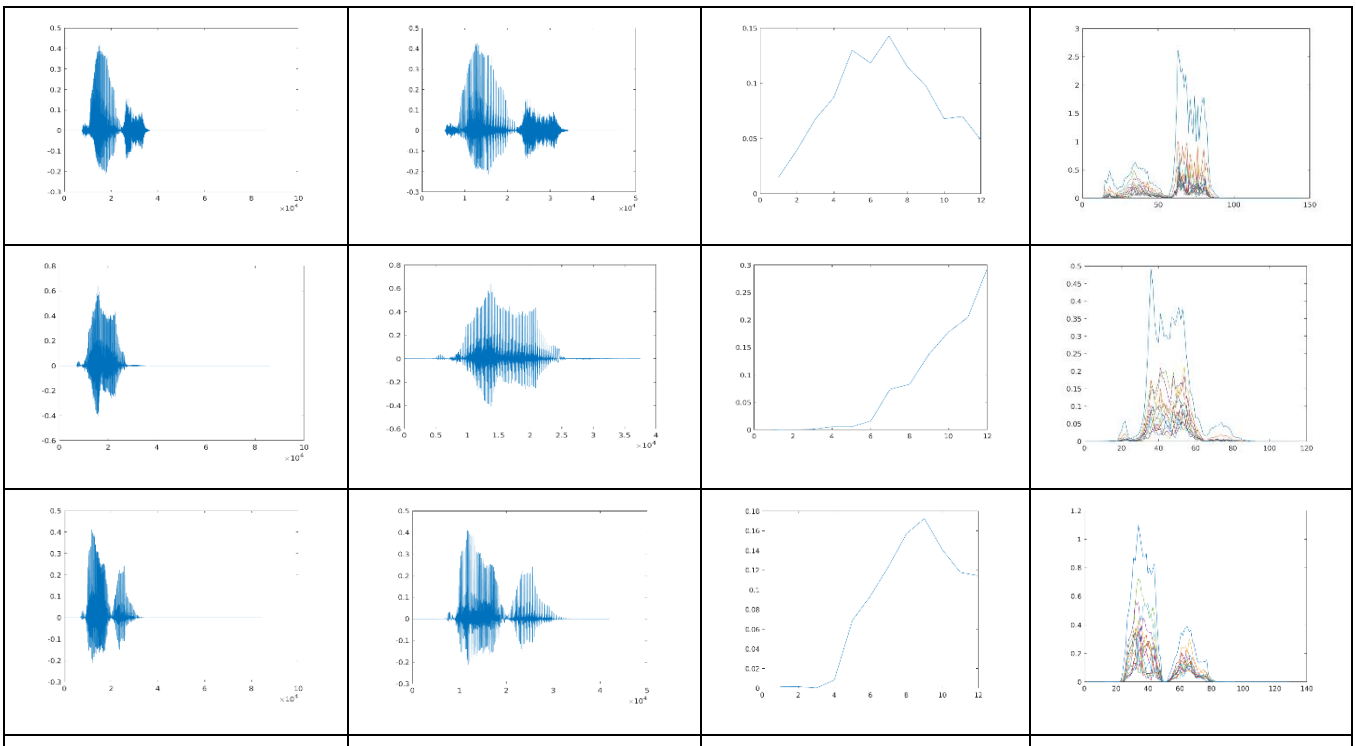
$$FPR = \frac{\mathrm{EP}}{\mathrm{TP} + \mathrm{EP}} \qquad (31)$$

d) Receiver Operating Characteristics (ROC) curve: ROC is a graphical demonstration of the TPR-FPR relationship, and it is the metric that defines the performance of the system.

d) Simulation analysis

Figure 3 shows the simulation results of the optimized DPM-based SBWOA. Figure 3a) refers to "Hash", "Health", and "Body" feedback signal. Figure 3b) displays the pre-processed "Hash" signal, "Health" signal, and "Body" signal, Figure 3c) shows the pitch chroma defined by the Hash, Health, and Body signal, and Figure 3d) demonstrates the AMS function obtained using the Hash signal, Health signal, and Body signal.

Figure 3 - Sample results of Hash, Health and Body signal of developed method a) Input signal, b) pre-processed signal c) pitch chroma signal, and d) AMS feature



a)	b)	c)	d)

e) Comparative methods

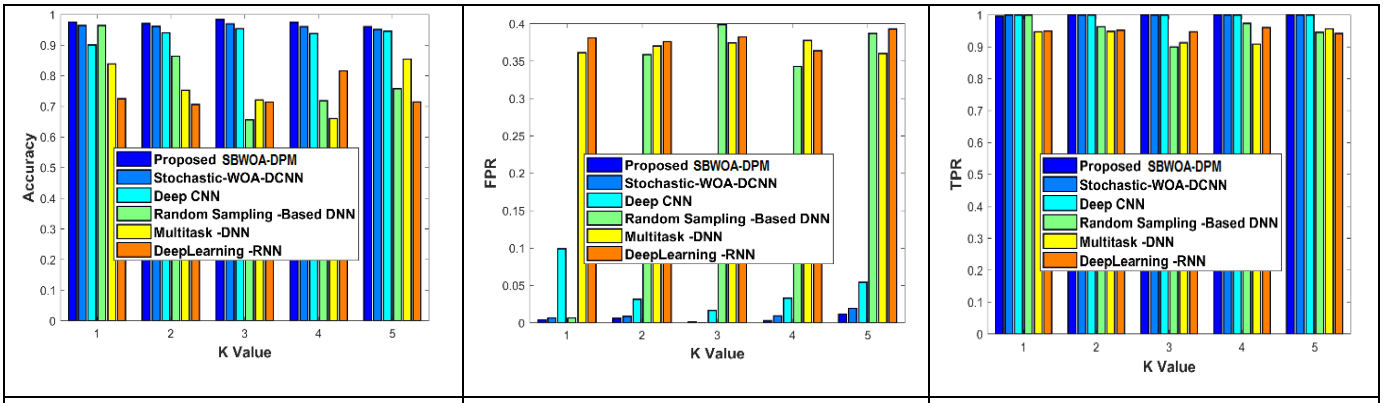
Compared with existing techniques, such as Stochastic-Whale Optimization Algorithm + Deep Convolutionary Neural Network (Stochastic-WOA+DCNN), Deep-CNN[31], Random sampling-based Deep Neural Network (DNN)[1], multi-task DNN acoustic models[4] and Deep

learning-Recurrent Neural Network (RNN)[8], the performance of the developed technique is evaluated.

a) Analysis using signal-1 based on K-Fold

The analysis of approaches using signal-1 using K-Fold by considering accuracy, FPR, and TPR parameters is illustrated in figure 4. The analysis of techniques based on accuracy parameter is deliberated in figure 4a. For K-Fold=5, the accuracy measured by Stochastic-WOA+DCNN is0.952, Deep-CNN is0.944, Random sampling-based DNN is0.756, multi-task DNN acoustic models is0.852, Deep learning-RNN is0.713, and the proposed SBWOA-DPM is 0.961. In Figure 4b, the analysis using the FPR parameter is deliberate. The FPR determined by Stochastic-WOA+DCNN is 0.02 when K-Fold=5, Deep-CNN is0.055, Random sampling-based DNN is0.388, multi-task DNN acoustic models are0.361, Deep learning-RNN is0.394, and the implied SBWOA-DPM is0.010. In Figure 4c, the analysis using the TPR parameter is intentional. When K-Fold=5, the TPR of methods computed by Stochastic-WOA+DCNN is 0.981, Deep-CNN is 0.973, Random sampling–based DNN is 0.944, multi-task DNN acoustic models is 0.955, Deep learning-RNN is 0.940, and the proposed SBWOA-DPM is0.998.

Figure 4 - Analysis of methods by signal-1 using K-Fold a) accuracy, b) FPR, and c) TPR



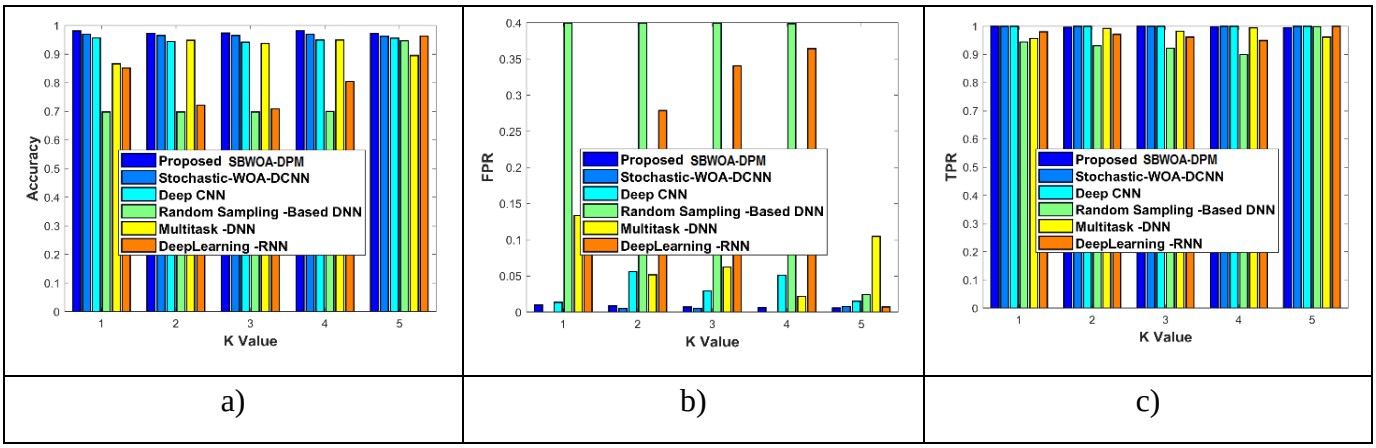
a) b) c)

b) Analysis based on signal-2 using K-Fold

The analysis of methods through signal-2by K-Fold by considering accuracy, FPR, and TPR parameters is illustrated in figure 5. The analysis of techniques by accuracy parameter is deliberated in figure 5a. When K-Fold=5, the accuracy computed by Stochastic-WOA+DCNN is 0.96, Deep-CNN is 0.948, Random sampling–based DNN is 0.698, multi-task DNN acoustic models is 0.947, Deep learning-RNN is 0.803, and the proposed SBWOA-DPM is0.982. The analysis using FPR

parameter is deliberated in figure 5b. When K-Fold=5, the FPR computed by Stochastic-WOA+DCNN is 0.005, Deep-CNN is 0.028, Random sampling–based DNN is 0.0232, multi-task DNN acoustic models is 0.131, Deep learning-RNN is 0.005, and the proposed SBWOA-DPM is 0.004. The analysis using TPR parameter is deliberated in figure 5c). When K-Fold=5, the TPR of methods computed by Stochastic-WOA+DCNN is 0.965, Deep-CNN is 0.962, Random sampling–based DNN is 0.89, multi-task DNN acoustic models is 0.984, Deep learning-RNN is 0.938, and the proposed SBWOA-DPM is0.998.

Figure 5 - Analysis of methods through signal-2 using K-Fold a) accuracy, b) FPR, and c) TPR

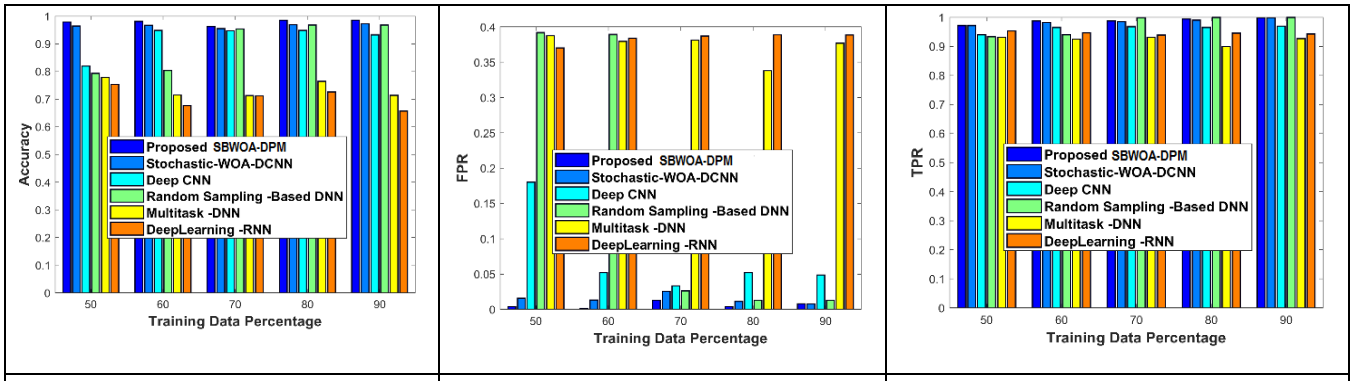


a) b) c)

c) Analysis through signal-1 by holdout

The analysis of methods through signal-1 using holdout by considering accuracy, FPR, and TPR parameters is illustrated in figure 6. The analysis of approaches based on accuracy parameter is deliberated in figure 6a. When training data percentage is 90, accuracy estimated by Stochastic-WOA+DCNN is 0.967, Deep-CNN is 0.937, Random sampling–based DNN is 0.964, multi-task DNN acoustic models is 0.714, Deep learning-RNN is 0.666, and the proposed SBWOA-DPM is0.984. The analysis using FPR parameter is deliberated in figure 6b. For 90% training data, FPR computed by Stochastic-WOA+DCNN is 0.014, Deep-CNN is 0.051, Random sampling–based DNN is 0.013, multi-task DNN acoustic models is0.378, Deep learning-RNN is 0.387, and the proposed SBWOA-DPM is0.013. The analysis using TPR parameter is deliberated in figure 6c). When training data percentage is 70, the TPR of methods computed by Stochastic-WOA+DCNN is 0.984, Deep-CNN is 0.968, Random sampling–based DNN is 0.999, multi-task DNN acoustic models is 0.930, Deep learning-RNN is 0.938, and the proposed SBWOA-DPM is0.938.

Figure 6 - Analysis of methods using signal-1 by varying the training data percentage a) accuracy, b) FPR, and c) TPR

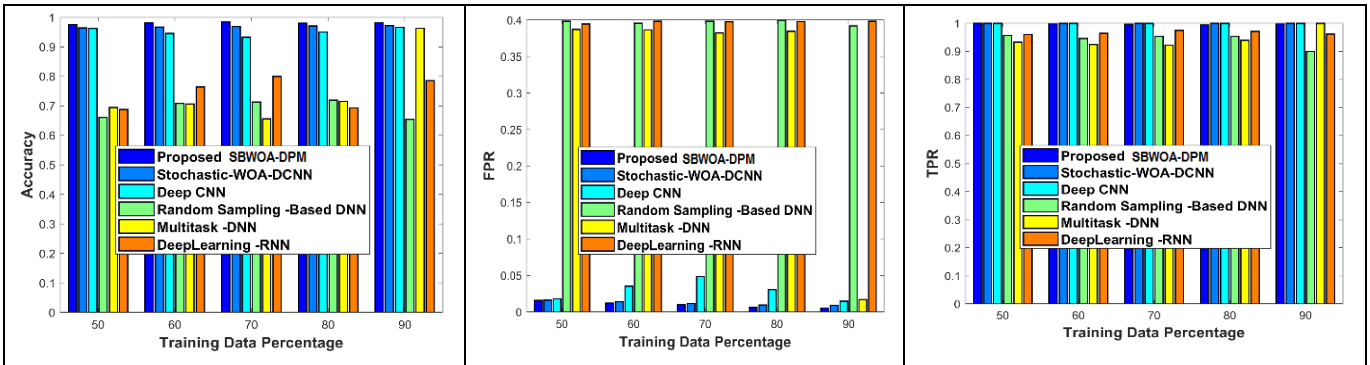


a) b) c)

d) Analysis by signal-2 using holdout

The analysis of methods based on signal-2 using holdout by considering accuracy, FPR, and TPR parameters is illustrated in figure 7. The analysis of approaches based on accuracy parameter is deliberated in figure 7a).When training data percentage is 70, accuracy calculated in Stochastic-WOA+DCNN is 0.967, Deep-CNN is 0.931, Random sampling–based DNN is 0.711, multi-task DNN acoustic models is 0.655, Deep learning-RNN is 0.9, and the proposed SBWOA-DPM is0.985. The analysis using FPR parameter is deliberated in figure 7b). When training data percentage is 80, FPR computed by means of Stochastic-WOA+DCNN is 0.010, Deep-CNN is 0.031, Random sampling–based DNN is 0.398, multi-task DNN acoustic models is 0.383, Deep learning-RNN is 0.399, and the proposed SBWOA-DPM is0.005. The analysis using TPR parameter is deliberated in figure 7c).When training data percentage is 70, the TPR of methods computed by Stochastic-WOA+DCNN is 0.993, Deep-CNN is 0.990, Random sampling–based DNN is 0.952, multi-task DNN acoustic models is 0.921, Deep learning-RNN is 0.973, and the proposed SBWOA-DPM is 0.996.

Figure 7 - Analysis of methods using signal-2 by varying the training data percentage a) accuracy, b) FPR, and c) TPR

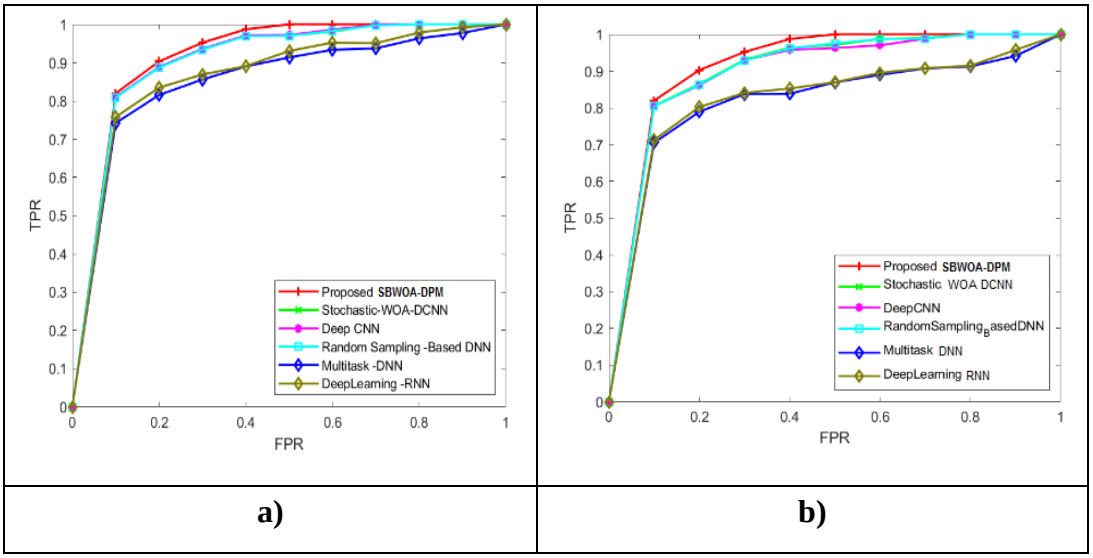


a) b) c)

f) ROC Analysis of signal-1 and 2

Figure 8 represents the ROC of the proposed SBWOA-DPM. The analysis of TPR vs. FPR of signal-1 is showed in figure 8a). The analysis of TPR is 0.887 for Stochastic-WOA+DCNN, 0.888 for Deep-CNN, 0.886 for Random sampling–based DNN, 0.815 for multi-task DNN acoustic models, 0.834 for Deep learning-RNN, and the proposed SBWOA-DPM is0.903, when the FPR is 0.2.The analysis of TPR vs. FPR of signal-2 is displayed in figure 8b). When FPR is 0.3, analysis of TPR is 0.930 for Stochastic-WOA+DCNN, 0.929 for Deep-CNN, 0.928 for Random sampling–based DNN, 0.837 for multi-task DNN acoustic models, 0.840 for Deep learning-RNN, and the proposed SBWOA-DPM is0.951.

Figure 8 - Analysis of methods of TPR vs. FPR a) Signal-1, and b) Signal-2



a) **b)**

g) Comparative discussion

Table 1 illustrates analysis of maximum performance attained by the methods by changing training data percentage using Hold out and K-Fold considering performance metrics. The maximal accuracy attained in Holdout, by the developed SBWOA-DPM with the value of 0.983, while accuracy of present Stochastic-WOA+DCNN is 0.971, Deep-CNN is 0.922, Random sampling–based DNN is 0.967, multi-task DNN acoustic models is 0.715, and Deep learning-RNN is 0.667. The minimal FPR computed by proposed SBWOA-DPM with a value of 0.005, whereas the FPR of existing Stochastic-WOA+DCNN is 0.009, Deep-CNN is 0.014, Random sampling–based DNN is 0.392, multi-task DNN acoustic model is 0.017, and Deep learning-RNN is 0.398. At last, the TPR

value achieved by proposed is 0.998, where as the current WOA-DCNN is 0.997, Deep-CNN is 0.969, RS based DNN is 0.968, Multi-task DNN acoustic model is 0.926 and Deep-RNN is 0.942.

In K-fold analysis, the proposed SBWOA-DPM achieved 0.984, whereas accuracy of present Stochastic-WOA+DCNN is 0.97, Deep-CNN is 0.953, Random sampling–based DNN is 0.656, multi-task DNN acoustic models is 0.721, and Deep learning-RNN is 0.714. The minimal FPR computed by proposed SBWOA-DPM with a value of 0.001, whereas the FPR of existing Stochastic-WOA+DCNN is 0.02, Deep-CNN is 0.016, Random sampling–based DNN is 0.398, multi-task DNN acoustic model is 0.374, and Deep learning-RNN is 0.382. In addition, the maximal TPR value measured by SBWOA-DPM is1, whereas the existing TPR of existing Stochastic-WOA+DCNN is 0.998, Deep-CNN is 0.997, Random sampling–based DNN is 0.921, multi-task DNN acoustic model is 0.981, and Deep learning-RNN is 0.960, respectively.

Table 1 - Comparative discussion of signal-1 and signal-2 using K-Fold and Holdout analysis

Variations	**Metrics**	**Stochastic-WOA+DCNN**	**Deep-CNN**	**Random sampling–based DNN**	**Multi-task DNN acoustic models**	**Deep learning-RNN**	**Proposed SBWOA-DPM**
Hold out	***Accuracy***	0.971	0.922	0.967	0.715	0.667	**0.983**
	FPR	0.009	0.014	0.392	0.017	0.398	**0.005**
	TPR	0.997	0.969	0.968	0.926	0.942	**0.998**
K-Fold	***Accuracy***	0.97	0.953	0.656	0.721	0.714	**0.984**
	FPR	0.02	0.016	0.398	0.374	0.382	**0.001**
	TPR	0.998	0.997	0.921	0.981	0.960	**1**

5. Conclusion

An enhanced form of speech recognition, called SBWOA-Deep CNN using a novel feature of Dirichlet Process Mixture, is proposed in this research work to convert non-audible murmurs to natural speech. Initially, the input signal is pre-processed using filtering to remove the noise in the signal. After that, the feature extraction is carried out based on Taylor AMS, pitch chroma, spectral centroid, spectral skewness, and proposed SBWOA-DPM in order to extract the appropriate features for further processing. Then, the speech recognition is carried out using extracted features using Deep CNN, which is trained using developed optimization algorithm, termed SBWOA. The incorporation of stochastic gradient descent methodology, WOA, and BBO is the SBWOA. Therefore, based on extracted functionality, developed SBWOA-based Deep CNN classifies speech signal as natural speech. Using the K-fold and holdout metric, the efficiency of the developed model is validated for

accuracy, FPR, and TPR. Comparatively, the proposed model dominated other methods with values of 0.984, 0.001, and 1 for accuracy, FPR, and TPR. In the future, speech prediction performance can be achieved through consideration of another algorithms and datasets for optimization.

References

Toktam Zoughi, Mohammad Mehdi Homayounpour, and Mahmood Deypir, "Adaptive windows multiple deep residual networks for speech recognition", *Expert Systems with Applications,* vol. 139, pp.112840, 2020.

Rongfeng Su, Xunying Liu, Lan Wang, and Jingzhou Yang, "Cross-Domain Deep Visual Feature Generation for Mandarin Audio-Visual Speech Recognition", *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol.28, pp.185-197, 2019.

Purvi Agrawal, and Sriram Ganapathy, "Modulation Filter Learning Using Deep Variational Networks for Robust Speech Recognition", *IEEE journal of selected topics in signal processing,* vol. 13, no. 2, May 2019.

Reza Yazdani, Jose-Maria Arnau, and Antonio Gonzalez, "LAWS: Locality-Aware Scheme for Automatic Speech Recognition", *IEEE Transactions on Computers,* 2020.

Thomas Hueber, and Gérard Bailly, "Statistical conversion of silent articulation into audible speech using full-covariance HMM", *Computer Speech and Language,* vol.36, pp.274-293, 2016.

D. Đorđe T. Grozdića, Slobodan T. Jovičića, and Miško Subotićb,"Whispered speech recognition using deep denoising autoencoder", *Engineering Applications of Artificial Intelligence,* vol.59, pp. 15–22, 2017.

Shabnam Ghaffarzadegan, Hynek Bořil, and John H. L. Hansen, "Deep neural network training for whispered speech recognition using small databases and generative model sampling", *International Journal of Speech Technology,* vol. 20, no. 4, pp. 1063-1075, 2017.

Đorđe T. Grozdic and Slobodan T. Jovicic, "Whispered Speech Recognition Using Deep Denoising Autoencoder and Inverse Filtering", *IEEE/ACM transactions on audio, speech, and language processing,* vol. 25, no. 12, December 2017.

Wentao Fan, Hassen Sallay, Nizar Bouguila, Sami Bourouis, "A hierarchical Dirichlet process mixture of generalized Dirichlet distributions for feature selection", *Computers and Electrical Engineering, Elsevier,* volume 43, PP.48–65, 2015.

Francois Caron, Manuel Davy, Arnaud Doucet, Emmanuel Duflos, Philippe Vanheeghe, "Bayesian Inference for Dynamic Models with Dirichlet Process Mixtures", *Proceedings of 9th IEEE International Conference on Information Fusion, Florence,* Italy. inria-00119993, 2006.

Slobodan T. Jovicic´ and Zoran Saric, "Acoustic Analysis of Consonants in Whispered Speech", *Journal of voice,* 22(3), pp.263-274, 2008.

Taisuke Ito, Kazuya Takeda, and Fumitada Itakura,"Analysis and recognition of whispered speech", *Speech Communication,* vol. 45, pp. 139–152, 2005.

Boon Pang Lim, "Computational differences between whispered and non-whispered speech (Doctoral dissertation, University of Illinois at Urbana-Champaign)", 2011.

Chen-Yu Yang, Georgina Brown, Liang Lu, Junichi Yamagishi1, Simon King, "Noise-robust whispered speech recognition using a non-audible-murmur microphone with vts compensation", *In 2012 8th International Symposium on Chinese Spoken Language Processing* (pp. 220-223). IEEE, 2012.

Arpit Mathur, Shankar M Reddy and Rajesh M Hegde, "Significance of parametric spectral ratio methods in detection and recognition of whispered speech", *EURASIP Journal on Advances in Signal Processing*, vol.1, pp.157, 2012.

Shabnam Gha. fJarzadegan, Hynek Bohl, and John H L. Hansen, "Generative modeling of pseudo-ta rget domain adaptation samples for whispered speech recognition", *In proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5024-5028). IEEE, 2015.

Szu-Chen Jou, Tanja Schultz, and Alex Waibel, "Adaptation for Soft Whisper Recognition Using a Throat Microphone", *In proceedings of Eighth International Conference on Spoken Language Processing*, 2004.

Fei Tao and Carlos Busso,"Lipreading Approach for Isolated Digits Recognition under Whisper and Neutral Speech", *In fifteenth annual conference of the international speech communication association*, 2014.

Healy EW, Yoho SE, Wang Y, and Wang D, “An algorithm to improve speech recognition in noise for hearing-impaired listeners", *Journal of Acoustical Society of America,* vol.134, no.4, pp.3029-38, October 2013.

Hermansky, Hynek, " Perceptual linear predictive (PLP) analysis of speech", *The Journal of the Acoustical Society of America,* vol.87, no.1738, 1990.

Fengbin Tu, Shouyi Yin, Peng Ouyang, Shibin Tang, Leibo Liu, and Shaojun Wei, "Deep Convolutional Neural Network Architecture with Reconfigurable Computation Patterns", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol.25, no.8, pp.2220-2233, 2017.

Ryan Sweke, Frederik Wilde, Johannes Jakob Meyer, Maria Schuld, Paul K. Fahrmann, Barthelemy Meynard-Piganeau, and Jens Eisert,"Stochastic gradient descent for hybrid quantum classical optimization", *Stochastic gradient descent for hybrid quantum-classical optimization. Quantum,* vol.4, pp.314, 2020.

Seyedali Mirjalili, and Andrew Lewis, "The Whale Optimization Algorithm", *Advances in Engineering Software*, vol.95, pp.51–67, 2016.

Dan Simon, "Biogeography-Based Optimization", *IEEE transactions on evolutionary computation,* vol. 12, no. 6, December 2008.

TIMIT Acoustic-Phonetic Continuous Speech Corpus, "https://catalog.ldc.upenn.edu/ldc93s1 ", 2018.

Rajesh Kumar T, Suresh GR, Kanaga Subaraja S, Karthikeyan C., “Taylor-AMS features and deep convolutional neural network for converting non-audible murmur to normal speech”, *Computational Intelligence,* Wiley Publisher, Pp.1–24. 2020.

Rajesh Kumar T., Lakshmi Sarvani Videla., Soubraylu SivaKumar, ASALG Gopala Gupta., D. Haritha, “Murmured Speech Recognition Using Hidden Markov Model”, *IEEE 7th International Conference on Smart Structures and Systems (ICSS)*, 2020.

Rajesh Kumar T., Dr. Suresh. G.R., S. Padmapriya., V. Thulasi Bai., P.M. Beulah Devamalar., “Conversion of Non-Audible Murmur to Normal Speech through Wi-Fi Transceiver for Speech

Recognition based on GMM Model", *IEEE sponsored second International Conference on Electronics and Communication Systems (ICECS),* 2015.

Panikos Heracleous, Yoshitaka Nakajima, Akinobu Lee, Hiroshi Saruwatari, Kiyohiro Shikano., "Non-Audible Murmur (NAM) speech recognition using a stetoscopic NAM microphone", 2004.

Diener, L., "The Impact of Audible Feedback on EMG-to-Speech Conversion", 2019.

T. Rajesh Kumar, G. R. Suresh, S. Kanaga Suba Raja, "Conversion of Non Audible Murmur to Normal Speech based on Full-rank Gaussian Mixture Model", *Journal of Computational and Theoretical NanoScience,* 1546-1955, Iss.1,vol-15, Pp:185-190, 2018.

Malaviya, H., Shah, J., Patel, M., Munshi, J. and Patil, H.A., "Mspec-Net: Multi-Domain Speech Conversion Network", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7764-7768). IEEE., 2020.

Rajesh Kumar T., Suresh GR and Kanaga Subaraja S., "Conversion of non-audible murmur to normal speech based on GR-GMM using Non-Parallal Training Adaptation Method", *Proceedings of ICISS-2019, IEEE Xplore,* 2019.

Tajiri, Y., Kameoka, H. and Toda, T., "A noise suppression method for body-conducted soft speech based on non-negative tensor factorization of air-and body-conducted signals"., *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4960-4964). IEEE., 2017.

Kumaresan, A., Mohankumar, N., Sureshanand, M. and Suganya, J., "Enhancing the Efficiency of Voice Controlled Wheelchairs Using NAM for Recognizing Partial Speech in Tamil". *Circuits and Systems,* vol. 7, no. 10, p.2884., 2016.

Shah, N.J., Parmar, M., Shah, N. and Patil, H.A., "Novel MMSE DiscoGAN for cross-domain whisper-to-speech conversion", *Machine Learning in Speech and Language Processing (MLSLP) Workshop,* pp. 1-3., 2018.

Chang, H.J., Liu, A.H., Lee, H.Y. and Lee, L.S., "End-to-end Whispered Speech Recognition with Frequency-weighted Approaches and Layer-wise Transfer Learning". arXiv preprint arXiv, 2020.

Shariff, M. N., Saisambasivarao, B., Vishvak, T., & Rajesh Kumar T., "Biometric user identity verification using speech recognition based on ANN/HMM", *Journal of Advanced Research in Dynamical and Control Systems,* 9(12 Special issue), pp.1739-1748. 2017.

Archana, V.G.S., "Embedded Sign communication Recognition Using KNN and HMM-VITERBI Fusion Classifiers", 2019.