Hindawi
Mobile Information Systems
Volume 2022, Article ID 8708380, 11 pages
https://doi.org/10.1155/2022/8708380

Research Article

Parallel Big Bang-Big Crunch-LSTM Approach for Developing a Marathi Speech Recognition System

Ashok Sharma,[1] **Ravindra Parshuram Bachate**,[2] **Parveen Singh**,[3] **Vinod Kumar**,[4] **Ravi Kant Kumar**,[5] **Amar Singh**,[6] **and Madan Kadariya**[7]

[1]*Department of Computer Science and Engineering, University of Jammu, Jammu, Jammu and Kashmir, India*
[2]*School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India*
[3]*Cluster University of Jammu, Jammu, Jammu and Kashmir, India*
[4]*Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India*
[5]*Department of Computer Science and Engineering, SRM University, Amaravati, AP, India*
[6]*School of Computer Applications, Lovely Professional University, Phagwara, Punjab 144001, India*
[7]*Department of IT Engineering, Nepal College of Information Technology (NCIT), Pokhara University, Lekhnath, Nepal*

Correspondence should be addressed to Madan Kadariya; madan.kadariya@ncit.edu.np

Received 16 June 2022; Revised 28 July 2022; Accepted 13 August 2022; Published 10 September 2022

Academic Editor: Praveen Kumar Reddy Maddikunta

The Voice User Interface (VUI) for human-computer interaction has received wide acceptance, due to which the systems for speech recognition in regional languages are now being developed, taking into account all of the dialects. Because of the limited availability of the speech corpus (SC) of regional languages for doing research, designing a speech recognition system is challenging. This contribution provides a Parallel Big Bang-Big Crunch (PB3C)-based mechanism to automatically evolve the optimal architecture of LSTM (Long Short-Term Memory). To decide the optimal architecture, we evolved a number of neurons and hidden layers of LSTM model. We validated the proposed approach on Marathi speech recognition system. In this research work, the performance comparisons of the proposed method are done with BBBC based LSTM and manually configured LSTM. The results indicate that the proposed approach is better than two other approaches.

1. Introduction

Due to the high degree of flexibility offered by speech recognition (SR) software and voice recording devices with multiple microphones, various models of hands-free speech communication are used in different types of application domain such as automatic speech recognition (ASR) and multimicrophone portables. Because of the effects of room reverberation, background noises, and interfering speakers on the considered speech signals, the performance of the automatic speech recognition model is generally minimized. Moreover, different speech enhancement techniques are intended to minimize the noise without affecting the speech signals to improve recognition models' robustness and performance [1]. However, automatic speech recognition models are complex due to constraints like freestyle or spontaneous speech, as well as lack of reliability to speech differentiations such as speaking rate, gender, sociolinguistics, accents, and environmental noise. There is a requirement for bridging the space among the speech recognition methods and humans to solve the challenges in these models.

Automatic speech recognition (ASR) models are crucial due to the complications present in the classification of languages with the common origin and intermixing of different languages along with the multilingual SC [2]. Therefore, there is a need to solve the limitations present in the existing recognition systems to get optimal results. Consequently, some of the research works have considered one of the South Asian languages like Marathi. However, there is no evidence for offering effective solutions while recognizing the Marathi language [3]. Moreover, the

Marathi language model suffers from inadequate SC and small size vocabulary systems [4].

Different deep learning approaches are well performed for SR fields. These approaches are used for automatic recognition models in "single-channel speech enhancement," and, thus, the recognition performance can be improved [5]. In existing studies, different speaker adaptation approaches are developed by targeting diverse speakers. Even though these existing deep learning algorithms often give more benefits, they also suffer from computational and language complexities [6]. After 12 kilometers of travel, the Marathi dialect is believed to shift. Due to numerous difficulties, speech signal processing is typically a challenging undertaking, yet effective research can provide solutions to all of these issues. Due to India's digitization, Marathi ASR and other Indian language ASR are in high demand [7]. Furthermore, the lack of research on Marathi language models inspires the researchers to design a new framework for Marathi language.

Significant contribution of the suggested framework is listed as follows:

(i) Developing novel framework on the Marathi language with multiple steps including preprocessing, feature extraction, and classification using a heuristic-based classification approach.

(ii) To extract the useful attributes from the speech signals with MFCC in addition to spectral-based features for increasing the performance. Here, the attained features are reduced to get significant features using Principal Component Analysis (PCA) technique.

(iii) Optimization of the hidden neurons and weight in the LSTM classification method using the PB3C algorithm to recognize speech signals to maximize the recognition accuracy.

The remaining sections of this paper discuss the literature survey and analyze the architectural view of Marathi SR, feature extraction, and feature selection for Marathi SR and the results and discussions. The paper ends with the conclusion.

2. Literature Survey

In 2021, Smit et al. [8] have described one new model for implementing subword language systems by considering the Deep Neural Networks (DNN), weighted finite-state transducers, and Hidden Markov Models (HMM). This paper has considered an acoustic system with character models and subword language systems without requiring the pronunciation dictionaries. They have also proposed approaches to combine the advantages of diverse classes of language model units through the reconstruction and combined recognition lattices. The developed model has constructed the Neural Network Language Models (NNLMs), which was practical due to fewer input and output layers. The four languages "Finnish, Swedish, Arabic, and English" were used to evaluate different subword units on SC. The experimental analysis was carried out and it showed more consistent results and reduced the error rate.

In 2019, Tu et al. [9] had developed a new Iterative Mask Estimation (IME) assembly for boosting the complex Gaussian mixture model- (CGMM-) based beamforming method to get the complete information. This model has developed a neural network- (NN) based ideal ratio mask estimator educated from the multicondition SC for incorporating the previous information. Subsequently, voice activity prediction information was attained from speech recognition results to use the rich context information in language models and deep acoustics, which was then employed to reduce the insertion errors and refine the mask estimation. The developed model experimented with the CHiME-4 Challenge ASR job of recognizing 6-channel microphone array speech in the testing process. The results of the experiments have revealed that the suggested IME method has consistently and significantly outperformed the existing CGMM method and reduced the error rate.

In 2017, Kipyatkova and Karpov [10] had implemented a Russian language automatic speech recognition model using recurrent artificial neural networks. It has considered hidden layers with different counting of elements, and the baseline trigram language model was performed with linear interpolation of NN models. The performance of the developed model was analyzed in terms of WER.

In 2015, Zhou et al. [11] had implemented a new "DNN-based acoustic modeling" structure for the ASR model, in which the multiple DNNs (mDNN) were computed to use the posterior probabilities of HMM states. Initially, the HMM states were clustered into different disjoint clusters by considering the data-driven approaches. Then, the mDNN was trained to cluster the states. They have shown that the considered training process using the mDNN model was employed to increase the training speed, including sequence-level discriminative training and frame-level cross-entropy. The suggested model has increased the capabilities of the developed model.

In 2014, Xue et al. [12] implemented a DNN-based ASR model by presenting different layers of pretrained DNN using a novel group of linking weights. Furthermore, the training approaches have learned a new condition code for each and every test condition from adaptation data. This developed model has used a fast adaptation strategy for developing an ASR model with supervised speaker adaptation. They have also implemented several speaker codes, in which the experimental analysis of the proposed adaptation scheme was carried out by comparison with different approaches. Lastly, they have attained superior performance in terms of WER, accuracy, and precision.

Bashir et al. [13] have proposed DNN-based emotion detection for Urdu language. The proposed DNN-based model outperforms other machine learning approaches. Akram et al. [14] projected a linguistic prototype for social text based on deep autoencoder. They have implemented this model for low resource language Urdu. The key addition in this exploration is converting high-dimensional feature space to low-dimensional one for Urdu language.

Table 1: Work done in the proposed area.

Author [citation]	Techniques used	Characteristics	Issue faced
Smit et al. [8]	RNN	It increases the performance with a better accuracy rate	This model is not suitable for a large amount of learning data
Tu et al. [9]	DNN and unidirectional LSTM	Reducing word error rate (WER)	This model is not suitable for executing the objective functions with joint learning
Kipyatkova and Karpov [10]	Artificial neural network	It reduces WER	However, this model suffers from the demographic influence on the languages
Zhou et al. [11]	mDNN	It increases the recognition performance and increases the training speed	Conversely, the accuracy rate can be degraded
Xue et al. [12]	DNN	(i) It improves the performance and efficiency while adapting larger DNN models (ii) It attains less WER	This model cannot optimize the speaker representations

2.1. Problem Statement. In recent years, different ASR models have been proposed, which are discussed in Table 1. RNN [8] increases the performance with a better accuracy rate. However, this model is not suitable for a large amount of training data. DNN and unidirectional LSTM [9] reduce the word error rate (WER). Conversely, it is not suitable for executing the objective functions with joint learning. Artificial neural network [10] reduces the WER. However, this model suffers from the demographic influence on the languages. mDNN [11] increases the recognition performance and increases the training speed. Conversely, the accuracy rate can be degraded. DNN [12] improves the performance and efficiency while adapting larger DNN models and attains less WER. On the other hand, this model cannot optimize the speaker representations. Moreover, the ASR model for the Marathi language is not focused on recent research works.

3. Architectural View of Marathi Speech Recognition (MSR)

From the past many years, more research studies have considered SR models using machine learning approaches. Different speech-related applications are focused on deep learning algorithms. Because of the usage of different ML and DL algorithms, speech recognition (SR) models are emerging areas in the research area.

3.1. Proposed Model and Description. Speech recognition models are crucial problems due to the complexities in determining local languages and correlation among different languages. Thus, the speech recognition framework on the Marathi language must be adopted with deep learning related approaches, represented in Figure 1.

Significant stages of the proposed framework for the Marathi language are "preprocessing, feature extraction, feature selection, and classification." Collected audio signals passed through preprocessing stage, which is done with smoothing and median filtering techniques. Moreover, the extracted features are reduced to get the optimal features using PCA to reduce the information's dimensionality. The selected attributes are forwarded to the labelling phase, in which the combination of LSTM with the PB3C algorithm

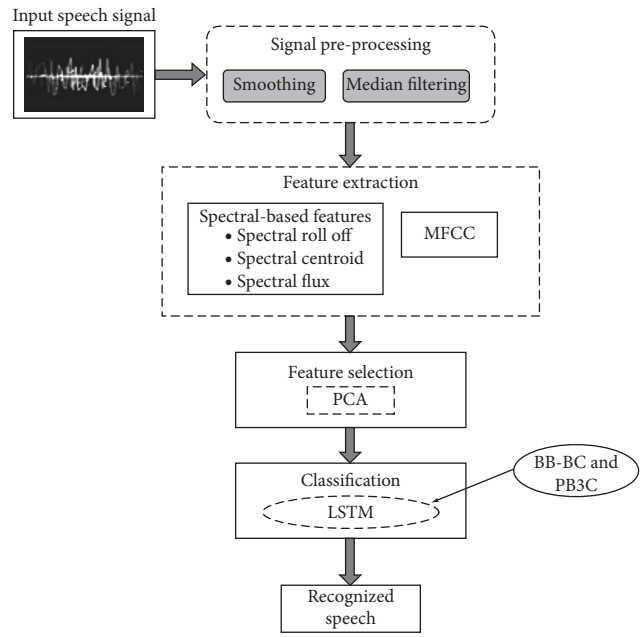

Figure 1: Architectural representation of the suggested framework.

takes place. Finally, the recognized speech signals are attained using the PB3C-based LSTM method. The P3BC optimization mechanism is applied to evolve the optimum quantity of neurons of different hidden layers of the LSTM. We also use PB3C algorithm to compute the weights of each link in the LSTM. This model aims to enhance the accuracy of LSTM model for Marathi speech recognition and find out light weighted machine learning model.

3.2. PB3C-LSTM Approach. The PCA-selected features are given to the PB3C-LSTM model for the efficient SR signals in the Marathi language. Here, the hidden neurons counting in LSTM is optimized with the assistance of PB3C technique. This model aims to improve the accuracy of speech-recognized signals.

In general, LSTM [15] is considered as the variants of the recurrent network by means of memory blocks. LSTM consists of input layer, hidden layers, and output layers. LSTM entails three gates, namely, output gate (OG), input gate (IG), and forget gate (FG). IG and OG are used to

regulate input and output functions in a block of memory cells. This is followed by the addition of the forget gate, where the LSTM network is used for getting the unit activations from the series of input Fs_n^{PCA}, where $n = 1, 2, \ldots, N$ and N stands for the number of features from PCA, which gets the output as $o_n = (o_1, o_2, \ldots, o_{N^{\text{optimal}}})$ to find mapping among them. It is equated as follows:

$$i_n = \alpha\left(wg_{im}Fs_n^{PCA} + wg_{iq}q_{n-1} + w_{ih}h_{n-1} + v_i\right), \quad (1)$$

$$f_n = \alpha\left(wg_{fm}Fs_n^{PCA} + wg_{fq}q_{n-1} + wg_{fh}h_{n-1} + v_f\right), \quad (2)$$

$$h_n = f_n \otimes h_{n-1} + i_n \otimes g\left(wg_{hm}Fs_n^{PCA} + wg_{hq}q_{n-1} + v_h\right), \quad (3)$$

$$p_n = \alpha\left(wg_{pm}Fs_n^{PCA} + wg_{pq}q_{n-1} + wg_{ph}h_{n-1} + v_p\right), \quad (4)$$

$$q_n = p_n \otimes jh_n, \quad (5)$$

$$o_n = \phi\left(wg_{pq}q_n + v_n\right). \quad (6)$$

In the above-mentioned equations, the FG bias vector (v_f), the OG bias vector (v_h), the input vector, or current time step is denoted as Fs_n^{PCA}, and the IG bias vector and the IG are termed as v_i and i, respectively. Moreover, the FG and the weight matrices are symbolized as fg and wg, respectively, and the cell activation vector, output gate, and the previous output from the blocks are denoted as p, h, and $p_{(n-1)}$, respectively. The cell output functions, the cell input functions, and sigmoid function are mentioned as j, go, and α, respectively, and ϕ is the output activation function. Further, the activation function (tanh) is employed in the multilayer LSTM. Similarly, the notations h_n and q_n indicate the memory of the current blocks and output of the current blocks, respectively. The peepholes connections diagonal weights are given as terms $wg_iFs_n^{\text{PCA}}$, $wg_fFs_n^{\text{PCA}}$, and $wg_pFs_n^{\text{PCA}}$, the highest weight value of the IG to the input is given as $wg_iFs_n^{\text{PCA}}$, and $i_{(n-1)}$ signifies the output coming of the preceding memory from input blocks. Here, the new LSTM is used by optimizing the number of hidden neurons using the PB3C algorithm. It is aimed at improving the classification accuracy to get accurate speech signals in the Marathi language.

3.2.1. BB-BC Algorithm. It is motivated by the "Big Bang Theory" in cosmological science that explains the conception of the cosmos in an explosion. The population is randomly distributed based on the center of mass (COM) in the big crunch stage. In the search space, the initial candidates are uniformly distributed in the BB-BC technique. Moreover, the big bang phase is immediately followed by the big crunch phase, where the fitness functions of each candidate and the current positions are replaced by the convergence operator for producing a weighted average point that is termed as a COM as formulated in the following equation:

$$c_{a,b}^{y} = \frac{\sum_{z=1}^{np}\left(1/fh_b^z\right)c_{a,b}^{z}}{\sum_{z=1}^{np}\left(1/fh_b^z\right)}. \quad (7)$$

In equation (7), the z^{th} solution of the fitness function at the b^{th} iteration is termed as fh_b^z, the a^{th} factor of the z^{th} answer at the b^{th} round is symbolized as $c_{a,b}^z$, the a^{th} factor of the center of a mass point at the b^{th} round is considered as $c_{a,b}^y$, the whole count of candidates in the population is shown as np, the current iteration is expressed as b, the current candidate in the population is denoted as z, and the current dimension is mentioned as a. The recent COM is considered as the essential in the next iteration and then explodes in the big bang phase. Further, the new members are produced by the explosion, which follows the normal distribution just about the COM as formulated here.

$$c_{a,(b+1)} = c_{a,b}^{z} + \frac{rn_a \times \delta \times \left(c_{\max} - c_{\min}\right)}{(b+1)}. \quad (8)$$

In equation (8), the standard normal distribution's random number is stated as rn_a, the upper and lower limits are termed as $c_{\max}$ and $c_{\min}$, respectively, parameter δ confines the parameters of the search domain, and the new candidate is noted as $c_{a,(b+1)}$. The optimal results are attained by fixing the value of standard deviation from equation (8), while the standard deviation is fixed for inversely decreasing the current iteration. The big crunch contraction phase is used for recalculating the COM, after the big bang explosion. Until the termination criterion is met, the explosion and contraction processes are continuously repeated. This BB-BC algorithm aims to attain the optimal results regarding SR in the Marathi language, which reflects the key goal as the maximization of recognition accuracy. Steps along with code of BB-BC algorithm are given in Algorithm 1.

3.2.2. PB3C Algorithm [17]. The extended version of BB-BC algorithm is a multipopulation optimization algorithm that shows superior accuracy and convergence rate while comparing with the BB-BC algorithm. This algorithm works by updating the elite by considering the local best solution in the population. The solutions are updated using the following equation:

$$c^{\longrightarrow y} = \frac{\sum_{z=1}^{np}\left(1/fh_b^z\right)c^{\longrightarrow z}}{\sum_{z=1}^{np}\left(1/fh^z\right)}. \quad (9)$$

The best fitness individual is selected based on the COM. Moreover, the new candidate solutions are updated around the COM as through subtracting or adding a normal random number that is decreased when the iterations are elapsed as given in equation (10). To generate the new population, we generate a change matrix between −1 and +1. The size of the change matrix should be similar to the extent of the candidate solutions in the population. We would get the new population after adding change matrix with the elite solution.

$$c_{\text{new}(z,k)} = \zeta_{\text{best}(z,k)} + \frac{\left(c_{\max} * rn\right)}{l}. \quad (10)$$

Here, the maximum number of iterations is termed as l, and a random number is mentioned as rn. PB3C algorithm is depicted in Algorithm 2.

```
Initialization of random number, population, and iteration
Estimate center of mass by equation (7)
While (b < B)
   Create new solutions by equation (8) around the center of mass
   Compute the fitness of every search agent
   Update new solutions
End while
Terminate
```

Algorithm 1: BB-BC algorithm [16].

```
Initialization of "N" population and each population consist of "C" candidate solutions
Compute the fitness of every candidate
Set i = 1
While (i < TC)
   for b = 1: N
      for j = 1 : C
         Compute the fitness of j^th candidate solution.
      end for
      Calculate the local best of b^th population
   end for
   Amongst the "N" local best solutions, find out the global best
   With the given probability, move the local best candidate solution towards the global best
   for b = 1: N
         Create new population around local best candidate solution
   end for
   i = i + 1
End while
Terminate
```

Algorithm 2: PB3C algorithm [18].

3.3. Objective Model. The suggested framework on Marathi language using P3BC-LSTM focuses on maximizing the accuracy to offer precise recognition. The objective function is formulated in the following equation:

$$fh = \underset{\{HN\}}{\arg\max}\,(Ac), \tag{11}$$

where fh represents the fitness function of the suggested SR model in Marathi language hidden neurons represented by HN. The accuracy is represented as Ac, which is an observation ratio to the whole observations as given in the following equation:

$$Ac = \frac{\left(po^{\text{true}} + po^{neg}\right)}{\left(po^{\text{true}} + po^{neg} + fa^{\text{true}} + fa^{neg}\right)}. \tag{12}$$

Here, po^{true} denotes true positive, fa^{true} denotes false positive, po^{neg} denotes true negative, and fa^{neg} denotes false negative.

4. Feature Extraction and Feature Selection of Marathi Speech Recognition (SR)

4.1. Signal Preprocessing. Signal preprocessing is the initial stage for processing. The collected signals are in an analog waveform that cannot be applied directly in any digital model. Smoothing and median filtering techniques have been used for preprocessing.

4.2. Smoothing [17]. It is used for reducing the noise present in the speech signals for noise reduction. A signal's data points are modified during the smoothing process, and the individual points are performed with their adjacent points, which have to be reduced in size as well.

4.3. Median Filtering [19]. It works in signal blocks. It aims at denoising the noise that existed in the input speech signal. The median filtering for the proposed Marathi speech signal is derived in the following equation:

$$\text{Median} = \begin{cases} ys\left(\dfrac{k-1}{2}\right), & K \text{ is odd}, \\ \dfrac{1}{2}\left[ys\left(\dfrac{k}{2}\right) + ys\left(\dfrac{k}{2}+1\right)\right], & K \text{ is even}. \end{cases} \tag{13}$$

A sorted set of K values is considered as $ys(k)$, while D is taken as odd. The term $ys(K - 1/2)$ indicates the middle

value of MF and the median filters are chosen while considering that the odd length is mentioned as K.

4.4. Feature Extraction. The preprocessed signals are fed to the feature extraction procedure to get the significant features using MFCC and spectral features.

4.4.1. MFCC [16]. MFCC is one of the significant feature extraction methods in the developed SR of Marathi language. It extracts all features, which are considered as given in the following equation:

$$CE_{mf} = \xi_{CE} \sum_{fp=0}^{FP-1} \cos\left(mf \frac{\pi}{FP}(fp + 0.5)\right) \log_{10}\left(En_{fp}\right). \tag{14}$$

In equation (14), ξ_{mf} represents the dynamic range coefficients, the amplification factor is represented as ξ_{CE}, and the energy in each channel is termed as En_{fp} as formulated in the following equation:

$$En_{fp} = \sum_{gw=0}^{GW-1} \sigma_{fp}(gw) Xs_{gw}. \tag{15}$$

Here, the value of fp lies among $0 \le fp < FP$, $G = 24$, and σ_{fp} denotes the number of triangular filters, where $Xs_{gw} = |\tilde{x}s(gw)|^2$ and $0 \le gw < GW$. Therefore, the features related to MFCC collected from the preprocessed signals are given to the next preprocessing step:

Spectral-based features [20]: the proposed model gathers spectral attributes like spectral (roll-off, flux, and centroid). The Fourier transform is used to convert the time-based signal into a frequency-based signal, which results in the appearance of spectral features. These techniques identify the pitch, notes, melody, and rhythm in the speech signals.

Spectral centroid: it is described as signal center of spectrum power distribution with distinct values for voiced and unvoiced speech. The sign function is derived as given in the following equation:

$$SpC = \frac{\sum_{nf}^{NF/2} fg(nf) Tu_r[nf]}{\sum_{nf}^{NF/2} |Tu_r[nf]|}. \tag{16}$$

Here, tr frame is mentioned as $Tu_r[nf]$, NF are Fourier transformation points, and a frequency is denoted as $fr(bn)$ at bin bn.

Roll-off is formulated in the following equation:

$$\sum_{nf}^{NF/2} |Tu_r[nf]| \le 0.85 \sum_{nf}^{NF/2} |Tu_r[nf]|. \tag{17}$$

Spectrum flux: the delta spectrum magnitude's Euclidean norm is derived in the following equation:

$$SF_r = \sum_{bn}^{NF/2} \left(|T_r[bn] - T_{r-1}[bn]|\right)^2. \tag{18}$$

The designed framework for recognition model excerpts the total number of attributes Fs_n^{MS} using both MFCC and spectral features as 100.

4.5. Feature Selection by PCA. The extracted features Fs_n^{MS} are given to this PCA [21, 22], which is used for concatenating the extracted features Fs_n^{MS} to get significant features Fs_n^{PCA} for further process. The PCA is computed by the following steps. Consider the data matrix as Fs_n^{PCA} with Vr variables and number of observations as Nu. Here, the PCA-based dimension reduction is derived in the following equation:

$$ps = Q' Fs_n^{MS}. \tag{19}$$

Here, the values of ps represent the principal components. The term Q is determined from the covariance matrix CM as derived in the following equation:

$$Q = Eg \cdot Dt^{-(1/2)}. \tag{20}$$

In equation (20), the matrix of eigenvectors of CM and diagonal matrix of the eigenvalues are termed as Eg and Dt, respectively. Assume that AC is the matrix of $Nu \times Vr$ with nu^{th} column as $Fs_{nu}^{MS} - \beta$.

$$AC = \left[Fs_1^{MS} - \beta, .., Fs_{nu}^{MS} - \beta\right]. \tag{21}$$

Here, the mean vector β is derived as $\beta = (1/Nu)(Fs_1^{MS} + ... + Fs_{No}^{MS})$ and CM is estimated with size of $Vr \times Vr$ as shown in the following equation:

$$CM = \frac{1}{No - 1} AC \cdot AC^T. \tag{22}$$

Finally, the attained PCA reduced features are denoted as Fs_{No}^{PCA}, where $No = 1, 2, \ldots, Np$ and Np is the total number of PCA reduced features, which is taken as 20.

5. Results and Discussion

This section explains the experimental setup, performance evaluation metrics taken for comparison of models, and the result analysis of proposed work.

5.1. Experimental Process. The developed framework takes into account a maximum of 25 iterations and a maximum of 10 populations in order to evaluate the performances. To evaluate algorithm's performance, it was tested on a Marathi SC obtained from the ILTPDC, Govt. of India, which was divided into six SCs for analysis. The collected SCs consisted of approximately 44500 speech files that were accompanied by their pronunciation. When using the LSTM model, the proposed PB3C-LSTM model was evaluated for performance [15] and BB-BC [23] on 6 SCs.

Table 2: Performance analysis of the designed framework on Marathi language with different algorithms for six different SCs in terms of error measures.

Speech corpus (SC)	Algorithms	WER	WAR	SER
SC 1	LSTM [21]	0.253333	0.826667	0.233934
	BBBC-LSTM [23]	0.233333	0.846667	0.221979
	PB3C-LSTM	0.193333	0.86	0.189498
SC 2	LSTM [21]	0.206667	0.86	0.190093
	BBBC-LSTM [23]	0.22	0.853333	0.206988
	PB3C-LSTM	0.206667	0.853333	0.231292
SC 3	LSTM [21]	0.266667	0.826667	0.26818
	BBBC-LSTM [23]	0.213333	0.866667	0.23345
	PB3C-LSTM	0.24	0.84	0.254593
SC 4	LSTM [21]	0.226667	0.853333	0.225522
	BBBC-LSTM [23]	0.226667	0.853333	0.230098
	PB3C-LSTM	0.22	0.853333	0.200145
SC 5	LSTM [21]	0.22	0.833333	0.223583
	BBBC-LSTM [23]	0.233333	0.826667	0.23908
	PB3C-LSTM	0.226667	0.866667	0.230612
SC 6	LSTM [21]	0.22	0.866667	0.215973
	BBBC-LSTM [23]	0.2	0.866667	0.216662
	PB3C-LSTM	0.18	0.88	0.194015

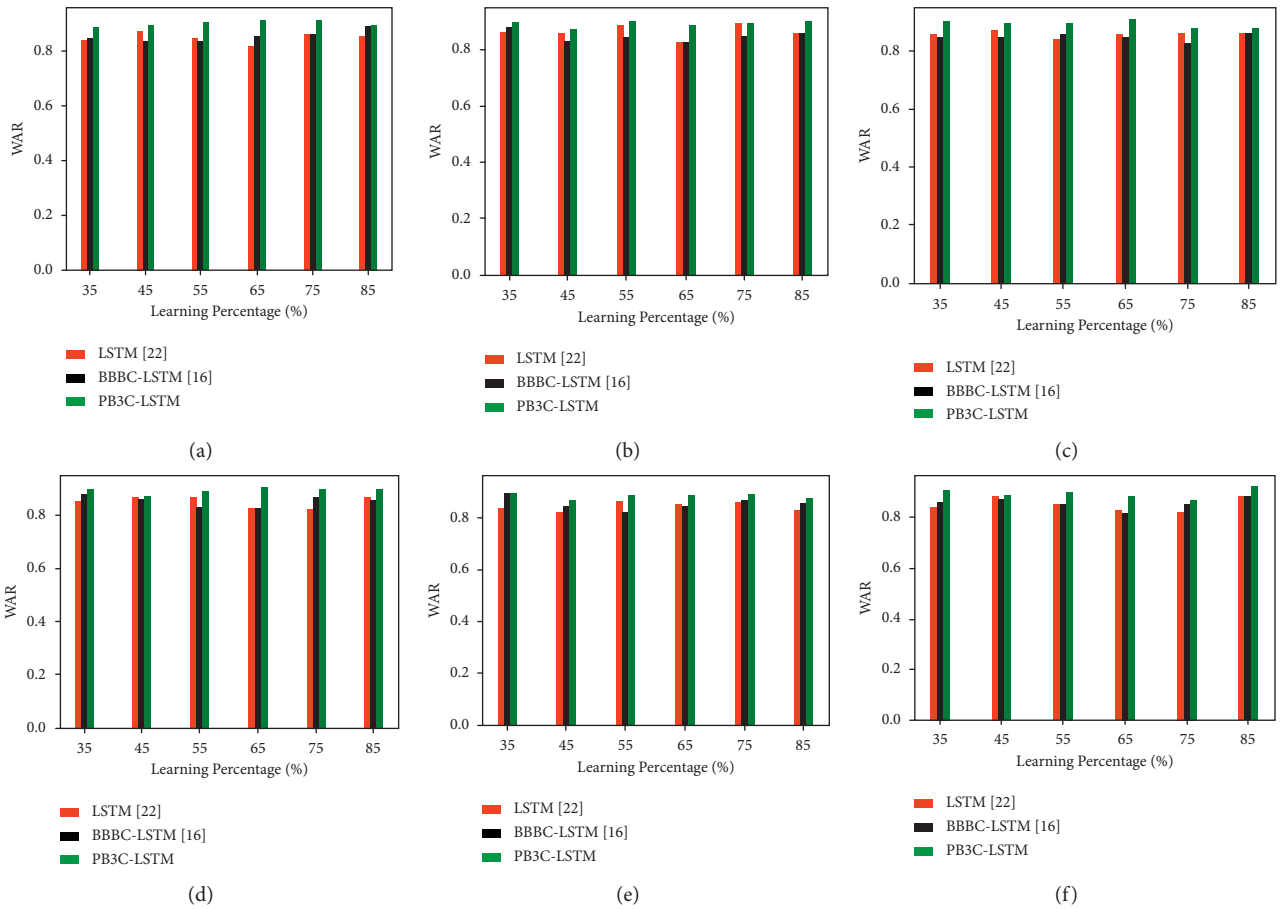

Figure 2: Word accuracy rate (WAR) analysis of the designed SR model on Marathi language with different conventional approaches for 6 different SCs in subgraphs (a–f).

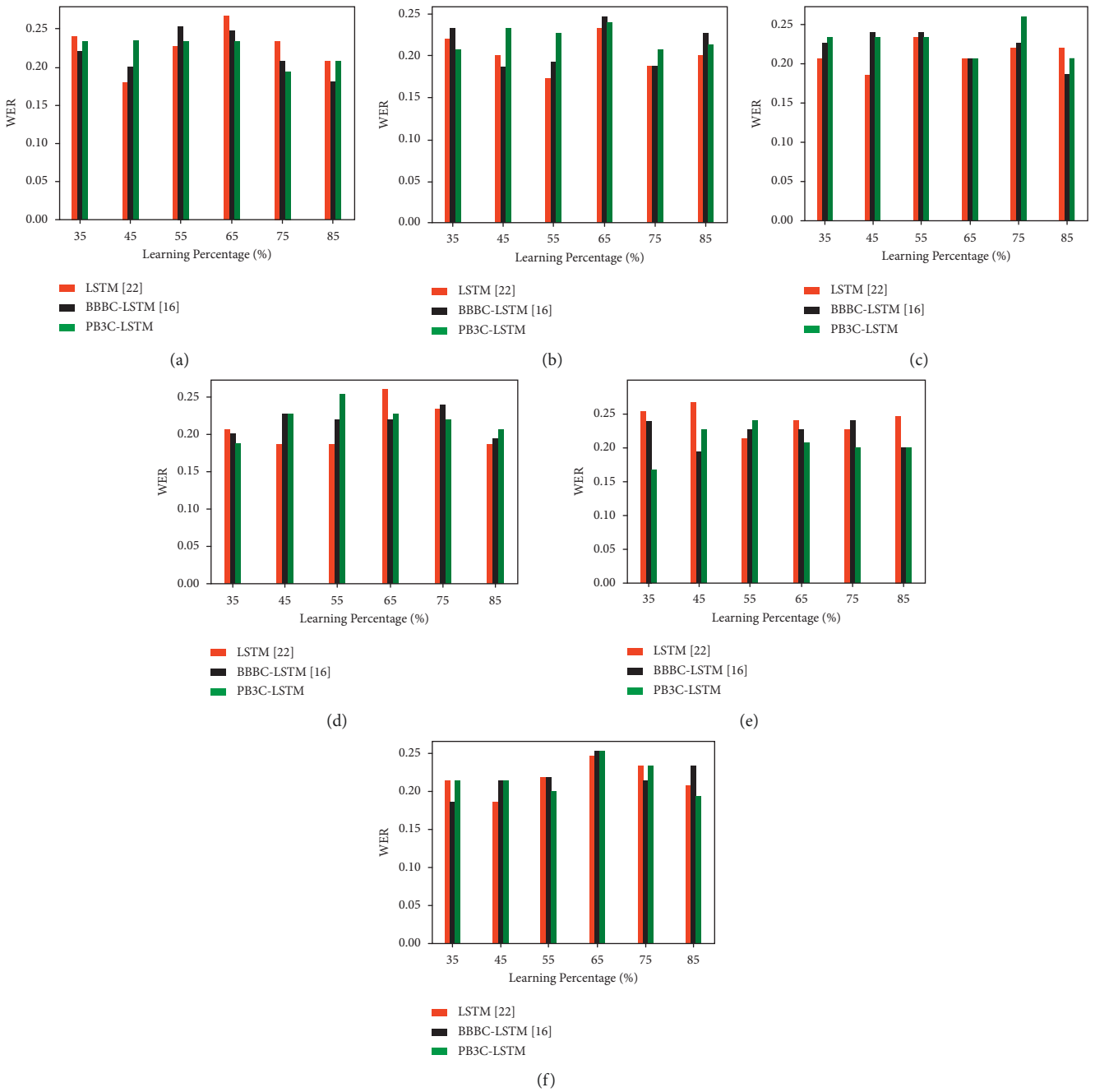

Figure 3: Word error rate (WER) analysis of the designed SR model on Marathi language with different conventional approaches for 6 different SCs in subgraphs (a–f).

5.2. Performance Measures. To evaluate the performance of LSTM, BB-BC LSTM, and PB3C-LSTM, word accuracy rate (WAR), WER, and sentence error rate (SER) are considered, which are described as follows:

(a) WER: it is used for measuring the word error rate of the designed framework.

$$\text{WER}(\%) = \frac{De + Ns + Ie}{Nw} * 100(\%). \tag{23}$$

Here, NS represents the number of substitutions in test, DT represents the number of deletions in the test, NT represents the number of words utilized in a test, and IE represents the number of insertion errors in the test.

(b) Word accuracy rate (WAR): it is used in measuring the word accuracy rate of the designed framework. It is derived in the following equation:

$$\text{WAR}(\%) = \frac{Nw - Ds - Ns}{Nw} * 100. \tag{24}$$

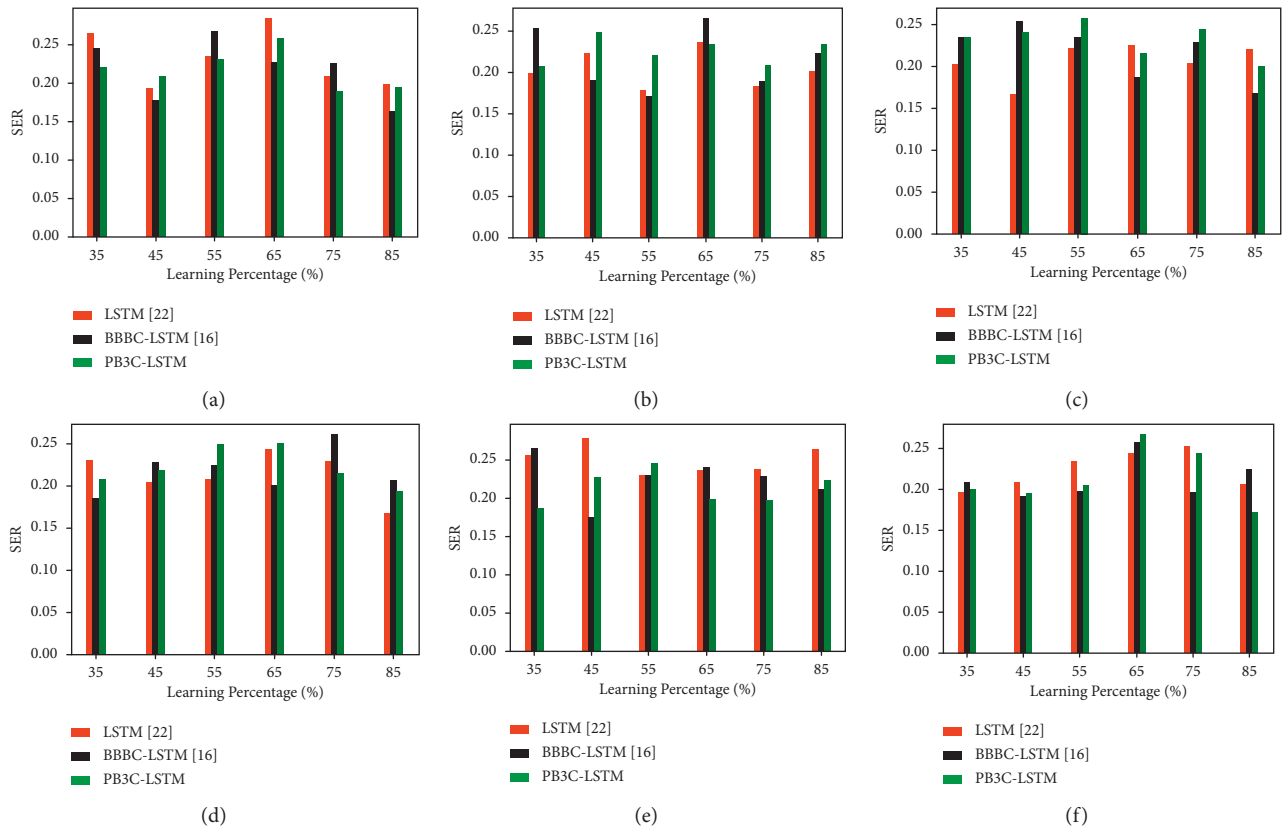

Figure 4: Sentence error rate (SER) analysis of the designed SR model on Marathi language with different conventional approaches for 6 different SCs in subgraphs (a–f).

(c) Sentence error rate (SER): it is correlated among audio predicted correctly to the total number of audios as given in the following equation:

$$\mathrm{SER}(\%) = 1 - \frac{PA}{TA} * 100. \tag{25}$$

Here, terms PA and TA represent the audio signals predicted correctly and the total number of audios, respectively.

5.3. Result Analysis. The results of the LSTM, BBBC-LSTM, and PB3C-LSTM SR model, are analyzed with error measures like WAR, WER, and SER as depicted in Table 2 and the results are presented in Figures 2–4, respectively. These measures are discussed in the following paragraphs.

Experimentation is conducted on WER for examining the efficiency of the designed PB3C-LSTM framework, given in Table 2 for all six SCs. The PB3C-LSTM is 6% and 4% boosted compared to LSTM and BBBC-LSTM, respectively. For SC 2, the PB3C-LSTM and LSTM have the same WER rate, but they are 1.33% superior to BBBC-LSTM. For SC 3, the WER performance of the PB3C-LSTM is 2.67% advanced and 2.67% declined compared to LSTM and BBBC-LSTM, respectively. For SC 4, the PB3C-LSTM is 0.67% and 0.67% progressed than LSTM and BBBC-LSTM, respectively. For SC 5, the WER performance of the PB3C-LSTM is 0.6% declined and 0.67% advanced compared to LSTM and BBBC-LSTM, respectively. For SC 6, the PB3C-LSTM is 0.4% and 0.6% boosted compared to LSTM and BBBC-LSTM respectively.

The WAR measure analysis is carried out to show the efficiency of the LSTM, BBBC-LSTM, and the proposed PB3C-LSTM SR model, which is given in Table 2 for all six SCs. For SC 1, PB3C-LSTM is 3.33% enhanced, and BBBC-LSTM is 1.33%, respectively. For SC 2, the WAR rate for PB3C-LSTM and BBBC-LSTM is 0.6 percent lower than that for LSTM. For SC 3, PB3C-LSTM's WAR performance is 1.33% advanced and 2.6% lower than the LSTM and BBBC-LSTM performance, respectively. The WAR of PB3C-LSTM is similar to that of the LSTM in terms of SC 4 and the BBBC-LSTM, respectively. For SC 5, the WAR performance of the PB3C-LSTM is 3.33% and 4% boosted than LSTM and BBBC-LSTM, respectively. The PB3C-LSTM for voice corpus 6 is 1.33% percent higher than the LSTM and the BBBC-LSTM, respectively.

The SER measure analysis for measuring the performances of the LSTM, BBBC-LSTM, and proposed PB3C-LSTM SR model is carried out, which is given in Table 2 for all six SCs. For SC 1, PB3C-LSTM is improved by 4.44% and 3.24% compared to LSTM and BBBC-LSTM, respectively. The performance of PB3C-LSTM decreased by 4.12% and 2.43% for SC 2, compared to LSTM and BBBC-LSTM. Speaking corpus 3 is 1.36% advanced and 2.1% lower than

LSTM and BBBC-LSTM performance in SER than LSTM and LSTM. The SER of PB3C-LSTM is up 2.53% and 2.99%, respectively, for SC 4 compared to the SER of LSTM and the BBBC-LSTM. SC 5 saw PB3C-SER LSTM's performance decrease by 0.7% and advance by 0.8% compared to LSTM and BBBC-LSTM. The PB3C-LSTM has been improved 2.19 percent and 2.26 percent, respectively, in the case of SC 6 compared to LSTM and BBBC-LSTM.

LSTM, BBBC-LSTM, and P3BC-LSTM had a mean WER performance percentage of 23.22%, 22.11%, and 21.11% correspondingly. The WAR mean LSTM, BBBC-LSTM, and P3BC-LSTM percentages are 84.44%, 85.22%, and 85.89%, respectively. The SER mean performance of LSTM, BBBC-LSTM, and P3BC-LSTM, respectively, is 22.62%, 22.47%, and 21.66%. The PB3C-LSTM was hence well functioning in comparison with LSTM and BBBC-LSTM techniques for Marathi SR System, taking into consideration all of the criteria.

6. Conclusion

This article has contributed new framework on Marathi language using P3BC-LSTM. It is composed of four significant processes: (1) classification, (2) feature selection, (3) feature extraction, and (4) preprocessing. The gathered speech signals were pretreated using smoothing and median filtering methods that were given to the feature extraction stage. It was carried out by MFCC and spectral-based attributes. Further, these attributes were significantly selected using PCA, which were forwarded to the classification stage using P3BC-LSTM. PB3C-LSTM module works in two phases. Phase 1 of PB3C-LSTM module automatically evolves the optimal architecture of LSTM. The second phase computes the optimum weight of each link in LSTM for Marathi SR. The P3BC-LSTM was proposed by optimizing the hidden neurons counting and optimized weights via the P3BC algorithm that intended to get the recognized speech signals. Consequently, from the experimental results, the word accuracy rate (WAR) of the proposed SR model using PB3C-LSTM was 1.34% and 3.34% increased compared to LSTM and BB-BC-LSTM, respectively, while considering SC 1. The proposed P3BC-LSTM model has attained 4% and 6% less WER and 4.45% and 3.25% less SER than LSTM and BB-BC-LSTM, respectively, for SC 1, and it has comparable performance with rest of the corpus. From results, we observe that the proposed system outperforms two other techniques for Marathi SR.

Data Availability

Marathi speech corpus is obtained from the ILTPDC, Govt. of India (URL: https://nplt.in/demo/resources/speech-corpus).

Conflicts of Interest

The authors declare that there are no conflicts of interest associated with the publication of this paper.

Acknowledgments

The authors express true sense of gratitude towards their fellow researchers and ILTPDC, Govt. of India, for providing Marathi speech corpus for this research.

References

[1] Y. H. Tu, J. Du, and C. H. Lee, "Speech enhancement based on teacher-student deep learning using improved speech presence probability for noise-robust speech recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.,* vol. 27, no. 12, pp. 2080–2091, 2019.

[2] L. M. Lee, "Adaptation of hidden Markov models for half frame rate observations," *Electronics Letters*, vol. 46, no. 10, pp. 723-724, 2010.

[3] G. S. V. S. Sivaram, S. K. Nemala, N. Mesgarani, and H. Hermansky, "Data-driven and feedback based spectro-temporal features for speech recognition," *IEEE Signal Processing Letters*, vol. 17, no. 11, pp. 957–960, 2010.

[4] L. Chai, J. Du, Q. F. Liu, and C. H. Lee, "A cross-entropy-guided measure (CEGM) for assessing speech recognition performance and optimizing dnn-based speech enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.,* vol. 29, pp. 106–117, 2021.

[5] S. Huang and S. Renals, "Hierarchical Bayesian language models for conversational speech recognition," *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, no. 8, pp. 1941–1954, 2010.

[6] V. Bhardwaj and V. Kukreja, "Effect of pitch enhancement in Punjabi children's speech recognition system under disparate acoustic conditions," *Applied Acoustics*, vol. 177, Article ID 107918, 2021.

[7] R. P. Bachate and A. Sharma, "Automatic speech recognition systems for regional languages in India," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2 Special Issue 3, pp. 585–592, 2019.

[8] P. Smit, S. Virpioja, and M. Kurimo, "Advances in subword-based HMM-DNN speech recognition across languages," *Computer Speech & Language*, vol. 66, Article ID 101158, 2021.

[9] Y. H. Tu, J. Du, L. Sun et al., "An iterative mask estimation approach to deep learning based multi-channel speech recognition," *Speech Communication*, vol. 106, no. 2018, pp. 31–43, 2019.

[10] I. S. Kipyatkova and A. A. Karpov, "A study of neural network Russian language models for automatic continuous speech recognition systems," *Automation and Remote Control*, vol. 78, no. 5, pp. 858–867, 2017.

[11] P. Zhou, H. Jiang, L. R. Dai, Y. Hu, and Q. F. Liu, "State-clustering based multiple deep neural networks modeling approach for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 631–642, 2015.

[12] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and L. Qingfeng, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," *IEEE/ACM Transaction Audio Speech Language Process*, vol. 22, no. 12, pp. 1713–1725, 2014.

[13] M. F. Bashir, A. R. Javed, M. U. Arshad, T. R. Gadekallu, W. Shahzad, and M. O. Beg, "Context aware emotion detection from low resource Urdu language using deep neural network," *ACM Transactions on Asian and Low-Resource Language Information Processing*, pp. 1–32, 2022.

[14] M. W. Akram, M. Salman, M. F. Bashir, S. M. S. Salman, T. R. Gadekallu, and A. R. Javed, "A novel deep auto-encoder based linguistics clustering model for social text," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2022.

[15] M. U. Abbasi, A. Rashad, A. Basalamah, and M. Tariq, "Detection of epilepsy seizures in neo-natal EEG using LSTM architecture," *IEEE Access*, vol. 7, pp. 179074–179085, 2019.

[16] R. Vergin, D. O'Shaughnessy, and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 525–532, 1999.

[17] M. A. J. Sathya and S. P. Victor, "Noise reduction techniques and algorithms for speech signal processing," *International Journal of Scientific Engineering and Research*, vol. 6, no. 1, 2015, http://www.ijser.org.

[18] S. Vaidya and D. K. Shah, "Audio denoising, recognition and retrieval by using feature vectors," *IOSR Journal of Computer Engineering*, vol. 16, no. 2, pp. 107–112, 2014.

[19] S. Herzog, "Efficient DSP implementation of median filtering for real-time audio noise reduction," in *Proceedings of the DAFx 2013-16th International Conference Digitital Audio Effect*, pp. 1–6, Maynooth, Ireland, December 2013.

[20] X. Kang, X. Xiang, S. Li, and J. A. Benediktsson, "PCA-based edge-preserving features for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 7140–7151, 2017.

[21] R. D. Aneja, A. K. Bindal, and S. Kumar, "HPGAB3C: a novel hybridized optimization approach," *Proceedings of Data Analytics and Management*, vol. 91, pp. 95–111, 2022.

[22] V. Kumar, "Evaluation of computationally intelligent techniques for breast cancer diagnosis," *Neural Computing & Applications*, vol. 33, no. 8, pp. 3195–3208, 2021.

[23] O. K. Erol and I. Eksin, "A new optimization method: big bang-big crunch," *Advances in Engineering Software*, vol. 37, no. 2, pp. 106–111, 2006.