Computer Systems Science & Engineering
DOI:10.32604/csse.2022.024214
Article

Tech Science Press

Enhanced Marathi Speech Recognition Facilitated by Grasshopper Optimisation-Based Recurrent Neural Network

Ravindra Parshuram Bachate[1], Ashok Sharma[2], Amar Singh[3], Ayman A. Aly[4], Abdulaziz H. Alghtani[4] and Dac-Nhuong Le[5,6,*]

[1]School of Computer Science and Engineering, Lovely Professional University, Punjab, 144001, India
[2]Department of Computer Science and IT, University of Jammu, Jammu and Kashmir, 180006, India
[3]School of Computer Applications, Lovely Professional University, Punjab, 144001, India
[4]Department of Mechanical Engineering, College of Engineering, Taif University, Taif, 21944, Saudi Arabia
[5]School of Computer Science, Duy Tan University, Danang, 550000, Vietnam
[5]Institute of Research and Development, Duy Tan University, Danang, 550000, Vietnam
*Corresponding Author: Dac-Nhuong Le. Email: ledacnhuong@duytan.edu.vn
Received: 09 October 2021; Accepted: 10 November 2021

Abstract: Communication is a significant part of being human and living in the world. Diverse kinds of languages and their variations are there; thus, one person can speak any language and cannot effectively communicate with one who speaks that language in a different accent. Numerous application fields such as education, mobility, smart systems, security, and health care systems utilize the speech or voice recognition models abundantly. Though, various studies are focused on the Arabic or Asian and English languages by ignoring other significant languages like Marathi that leads to the broader research motivations in regional languages. It is necessary to understand the speech recognition field, in which the major concentrated stages are feature extraction and classification. This paper emphasis developing a Speech Recognition model for the Marathi language by optimizing Recurrent Neural Network (RNN). Here, the preprocessing of the input signal is performed by smoothing and median filtering. After preprocessing the feature extraction is carried out using MFCC and Spectral features to get precise features from the input Marathi Speech corpus. The optimized RNN classifier is used for speech recognition after completing the feature extraction task, where the optimization of hidden neurons in RNN is performed by the Grasshopper Optimization Algorithm (GOA). Finally, the comparison with the conventional techniques has shown that the proposed model outperforms most competing models on a benchmark dataset.

Keywords: Deep learning; grasshopper optimization algorithm; recurrent neural network; speech recognition; word error rate

1 Introduction

Speech recognition is an emerging and attractive research area in today's world [1]. The recognition of speech in Indian regional languages is a broadly concerned field because of the complexities in such languages. The advancements in technologies have emerged over the country without the bondage of language, which can be realized using the speech recognition model with the regional languages [2]. Speech recognition is generally continued by the vocabulary that insists on operating the regularly utilized words. The term vocabulary includes words that must be treated in linguistics. Marathi is an Indo-Aryan Language that is phonetic in nature and spoken in central and western India. There is a major scope for developing the Speech Recognition models with Indian languages presented in various distinctions [3]. In Indian regional languages, the real work has not yet been accomplished to a significant level by using a simple communication tool like any other language. Therefore, the research has been considered for focusing on the Marathi language [4]. The primary aim of this paper is to establish the implementation of the Marathi Speech Recognition system.

A Speech Recognition model is accurately trained through the annotated transcriptions of speech [3]. The quantity of actual annotated data stored in reverberant and noisy conditions is very much restricted, especially when contrasted to the amount of data that can be experimented with the addition of noise for cleaning annotated speech [5]. Therefore, both simulated and real data are used significantly for improving robust speech recognition for increasing the diversity and amount of training data [6].

Additionally, multi-task learning is applied in the speech recognition model on reverberant and noisy conditions. The training of an acoustic system is intended for solving two or more tasks simultaneously, which are different in nature. The purpose of the auxiliary task includes generating the features of clean speech through a regression loss [7].

The most efficient deep learning techniques used for developing Speech Recognition Systems are based on the advancements in supervised learning [4], which are applied to classify the tasks. It relies on the annotation of the data that may be resource-intensive and time-consuming [8]. Due to the non-reverberant and clean nature of the acoustic environment, the amount of offered annotated data is extremely significant for the Speech Recognition System [3]. Though, these perfect acoustic conditions are not much reasonable as in diverse many real-life conditions and may be suffered by the deprivations of the speech signal that is come from the room proprieties of the acoustic or from the noise in the surrounding, which directs to the reverberations of the speech [7]. These occurrences can formulate the speech recognition model that attains significantly depreciate results and is a much more challenging task. The other difficulty in this reverberant and noisy case is the restriction of annotated actual data [9].

The purpose of this paper is to deal with the issues in developing the Marathi Speech Recognition system. The major issue with the Marathi Speech recognition is that available Marathi Speech Recognition Systems have not considered the dialects of the Marathi language and only considered the Marathi language dialect which is spoken in a small part of the state. The significant contribution of this proposed speech recognition model is listed here.

- To develop a speech recognition model on Marathi language using optimized RNN with GOA by diverse stages like preprocessing, feature extraction, and classification.
- To extract the features such as MFCC and spectral-based features like spectral rod off, spectral centroid, and spectral flux in the feature extraction phase from the preprocessed signal.
- To optimize the number of hidden neurons in the RNN classifier using GOA for the classification of extracted features to obtain the recognized speech.

The remaining sections of this paper are listed here. Section 2 discusses the literature survey. Section 3 explains the development of the proposed speech recognition model. Section 4 describes the feature

extraction and optimized RNN adaptable for the proposed speech recognition model. Section 5 shows the results and discussions. Section 6 concludes this paper.

2 Related Work

In 2020, Pironkov et al. [10] had proposed an HTL model. It often switched among single and multi-task learning by considering the input as either simulated or real data, respectively. Hybrid architecture has allowed by benefiting from both simulated and real data when employing a denoising auto-encoder like an auxiliary task of the setup of a multi-task. This HTL architecture has improved the performance of the conventional single-task learning method on the "CHiME4" database.

In 2010, Sivaram et al. [11] implemented new feedback and data-driven-based "discriminative spectro-temporal filters" to extract the features in ASR. Firstly, a set of "spectro-temporal filters" were planned for separating every phoneme from the remaining phonemes. A hybrid HMM/MLP phoneme recognition model was used for training the features obtained through such filters. The confusions were addressed specifically for designing the second set of the filter in the identification of top confusions in the model. From the experimental results, phoneme recognition has established better features and contained important corresponding information than conventional speech recognition models.

In 2020, Guglani et al. [12] had proposed the development in the performance of the ASR model using the possibility of voicing computed features and pitch-dependent features. The ASR model with tonal language as Punjabi was performed using the pitch-dependent features. Therefore, the ASR model was built using the possibility of voicing computed features and pitch-dependent features for the Punjabi language. The performance of the ASR model was measured using the WER measure that has significantly improved the features. The features like Kaldi pitch, FFV, SAcC, and Yin were performed concerning the WER. The proposed Kaldi toolkit was used for achieving improved performance among the other featured ASR models. In 2010, Huang and Renals [1] exploited an HPYLM using a hierarchical Bayesian model. It has presented a principled technique for embedding the power-law distribution, smoothing of a language model for natural language. The recognition of the conversational speech was experimented with by hierarchical Bayesian language models, which was achieved by substantial reductions in WER and perplexity. Thus, the convergence of HPYLM is a significant aspect.

In 2021, Smit et al. [13] have employed diverse approaches and a set of tools to successful subword modeling for WFST-based hybrid DNN-HMM speech recognition by graphemes. This model has also evaluated the four different languages and estimated such approaches in an under-resourced situation. Furthermore, the subword usage and considerations were explored by character-based NN-LM for hybrid DNN-HMM models. Finally, the evaluation of these tools over various language modeling units was done.

The GOA-RNN [14] has proposed to design and optimize the controller for wind energy hybrid-fed pumping systems. The GOA optimizes a resource's parameters based on solar radiation and wind power uncertainty. The RNN generates the best control signals, such as duty cycle, based on the best datasets. Using the GOA algorithm in the context of the minimum error objective function improves the RNN learning process.

Problem Statement

Diverse speech recognition models have observed various significant developments like Cortana, Google Assistant, Siri, etc. The Speech Recognition models should be constructed for regional languages as most populations in India are not well-known for English. However, there is less number of ASR systems with the Indian regional languages than the English language. Therefore, there is a necessity for developing the ASR model to adopt the current service sector. However, there are diverse limitations observed in the development of the ASR model; one among them is noise reduction. Varied speech

recognition models are introduced in the previous year with various restrictions and features, as given in Tab. 1.

Table 1: Features and challenges of conventional speech recognition models

Authors	Methodology	Features	Challenges
Pironkov et al. [10]	HTL	It improves performance and flexibility, and It can be applied to various kinds of dataset.	This method is limited due to the convergence of the auxiliary task.
Sivaram et al. [11]	HMM/MLP	It attains better performance. It can easily capture additional features.	It discriminates the spectro-temporal patterns.
Guglani and Mishra [12]	Pitch dependent features	It reduces WER and improves speech recognition.	It observes the variation of frequencies in the experimentation.
Huang and Renals [1]	Hierarchical Bayesian model	It improves accuracy. The performance of the model is enhanced.	However, this model has the issue of smoothing.
Smit et al. [13]	WFST	The robustness of this model is increased. It reduces the sparsity of this model.	This model does not apply to the integration of graphene and phoneme-based systems.

HTL [10] is proposed for improving performance and flexibility. It can be applied to several kinds of the dataset. Though, this method is limited due to the convergence of the auxiliary task. HMM/MLP [11] is developed for attaining better performance and easily captures the additional features. However, it discriminates the spectro-temporal patterns. Pitch-dependent features [12] reduce the WER and improve speech recognition. Anyhow, it observes the variation of frequencies in the experimentation. The hierarchical Bayesian model [15] is proposed for improves accuracy and performance. However, this model has the issue of smoothing. WFST [13] enhances robustness and reduces sparsity. On the other hand, this model does not apply to the integration of graphene and phoneme-based systems. These limitations are considered for developing a new speech recognition model.

3 Development of Proposed Speech Recognition Model

3.1 Proposed Architecture

Implementing the Speech Recognition model for Indian languages is complex and enables broad motivation and growing research in this area. Most of the deep learning algorithms are still based on supervised learning in the area of implementing a Speech Recognition System. However, supervised learning relies on the annotation of the data that is a resource-intensive and time-consuming approach. The non-reverberant and clean acoustic atmosphere produce a better Speech Recognition model. Though such ideal acoustic cases are not much realistic as in diverse real-life circumstances, and hence, the Speech Recognition model suffers from the degradations of the speech signal. Implementing the Speech Recognition model for different languages requires numerous techniques due to the nature of each language. It has its phonetical utterances and grammar collection. For Indian regional languages like Marathi, the Speech Recognition model is a broader research area. The major challenge in developing the speech recognition model is the reduction of noise in audio signals that affects the performance of the

proposed model. Therefore, there is a necessity to establish an efficient speech recognition model for the Marathi language shown in Fig. 1.

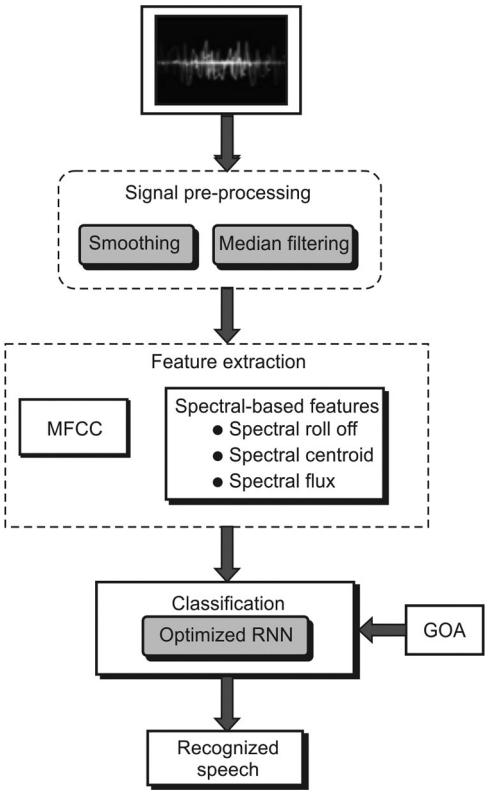

Figure 1: Proposed speech recognition model on the marathi language

The proposed speech recognition model on the Marathi language includes three phases preprocessing, feature extraction, and classification. The audio signal is given as an input to the proposed speech recognition model. In the initial stage, the preprocessing is performed using the smoothing and median filtering approaches. Preprocessing is the process of eliminating the artifacts and the noise from the signal. Secondly, the feature extraction process is carried out for extracting the features MFCC and spectral-based features such as spectral rod off, spectral centroid, and spectral flux. The process of feature extraction is done for discarding the inappropriate information but keeping the useful one. The extracted features are subjected to the classification phase, in which the training is executed with those extracted features from the input signal. Here, the proposed speech recognition model on the Marathi language uses optimized RNN for the classification of features, where the hidden neurons of RNN are optimized using the optimization algorithm called GOA. The proposed model has attained better accuracy and minimum WER in making the efficient speech recognition model on the Marathi language.

3.2 Preprocessing of Input Signals

This is an initial stage of the proposed speech recognition model, where the input is considered as the speech signal. This signal is an analog waveform that cannot be straightly used with digital models. Therefore, the preprocessing is performed for transforming the input speech into a form, which can be recognized using the recognizer. The proposed speech recognition model utilizes the two methods such as smoothing and median filtering that are discussed below.

Smoothing: It is a crucial noise reduction approach that is used in the preprocessing stage for the proposed speech recognition model. It removes the noise in the audio signal. The data points of a signal are altered in smoothing; thus, individual points are superior to the adjacent points that are reduced. The points which are lesser than the adjacent points are maximized, which leads to a smoother signal.

Median filtering [16]: It is a popular and well-known signal processing block, which is majorly used for denoising by reducing some kind of noise in the input signal. The median filtering is formulated in Eq. (1).

$$MF = \begin{cases} y\left(\dfrac{D-1}{2}\right) & D\ odd \\ \dfrac{1}{2}\left[y\left(\dfrac{D}{2}\right) + y\left(\dfrac{D}{2}+1\right)\right] & D\ even \end{cases} \tag{1}$$

Here, the term MF denotes the median of a set of values that is separated into two parts. Consider a sorted set of values of D as $y(d)$, where D is odd. The simple element of MF middle value is considered as $y\left(\frac{D-1}{2}\right)$. Generally, median filters are selected for having the odd length as D.

4 Feature Extraction and Optimized RNN Adaptable for Proposed Speech Recognition Model

4.1 Feature Extraction Process

It is the second stage of the proposed speech recognition model on the Marathi language for attaining diverse features. The key aim of feature extraction is to discover a set of properties of an utterance, which are called features. It is the process of getting significant information by avoiding irrelevant ones [17]. This process consists of computing various significant characteristics of the signal like frequency response or energy. A sequence of sounds is called speech that has various properties related to it, which are converted into tiny parts called a frame. Thus, these frames are processed for extracting the significant features concerned with speech vectors. The proposed model extracts the features such as MFCC and spectral-based features such as spectral rod off, spectral centroid, and spectral flux.

MFCC [18]: One of the significant features of the speech recognition model is MFCC. It is derived from the human input speech signal. It is "a representation of linear cosine transforms of a short-term log power spectrum of a speech signal on a non-linear Mel scale of frequency". Extraction of MFCC is of the type in which entire characteristics of the speech signal are focused in the initial few coefficients. MFCC is computed in Eq. (2).

$$Co_m = \delta_{Co} \sum_{f=0}^{F-1} \cos\left(m \frac{\pi}{F}(f+0.5)\right) \log_{10}(En_f) \tag{2}$$

In Eq. (2), the term δ_{Co} referred to as the amplification factor, which Co_m denotes the dynamic range of the coefficients that is dependent on the normalization factor, and the term En_f shows the energy in each channel, which is given in Eq. (3).

$$En_f = \sum_{g=0}^{G-1} \sigma_f(g) X_g \tag{3}$$

Here, $0 \leq f < F$, $G = 24$, the number of triangular filters used is considered as σ_f, where $X_g = |\tilde{x}(g)|^2$ and $0 \leq g < G$. Thus, the MFCC features collected from the input signal are further subjected to the classification stage along with the spectral-based features.

Spectral-based features [19]: It is also called frequency-based features that are attained by switching the time-based signal into the frequency domain by Fourier transform. It is employed for identifying the pitch,

notes, melody, and rhythm. Some of the spectral-based features are spectral roll-off, fundamental frequency, spectral flux, spectral density, frequency components, and spectral centroid. Here, the proposed speech recognition model extracts the spectral centroid, spectral flux, and spectral roll-off that are explained below.

Spectral centroid: It is defined as "the center of a signal's spectrum power distribution. It has different values for voiced and unvoiced speech". It is the gravity of the spectrum, in which the sign function is given in Eq. (4).

$$SC = \frac{\sum_{bn}^{NF/2} fr(bn) T_r[bn]}{\sum_{bn}^{NF/2} |T_r[bn]|} \tag{4}$$

Here, the STFT of frame of *tr* is given as $T_r[bn]$, the number of FFT points is specified as *NF* and a frequency at bin *bn* is termed as $fr(bn)$.

Spectral roll-off: It is "the measure of skewness of the signal's frequency spectrum and the frequency below which 85% of the magnitude distribution of the spectrum is concentrated is known as Roll-Off" that is given in Eq. (5).

$$\sum_{bn}^{NF/2} |T_r[bn]| \leq 0.85 \sum_{bn}^{NF/2} |T_r[bn]| \tag{5}$$

Spectrum flux: It is "a measure which characterizes the change in the shape of the signal's spectrum. It is calculated as the ordinary Euclidean norm of the delta spectrum magnitude" that is specified in Eq. (6).

$$SF_r = \sum_{bn}^{NF/2} (|T_r[bn] - T_{r-1}[bn]|)^2 \tag{6}$$

The proposed speech recognition model extracts the total number of features as 15.

4.2 Recurrent Neural Network

RNN [20–25] is an efficient method for the speech recognition model because of the 'end-to-end'. structure for sequential data. RNN efficiently gives its perspective on speech recognition. However, the performance of RNN in speech recognition is disappointing, as given in the literature. Thus, this proposed model utilizes the optimized RNN for improving performance. The RNN classifier is enhanced by the renowned optimization algorithm called GOA. RNN is intended for sequence-to-sequence mapping or prediction. Assume the input sequence as $x = (x_1, \ldots, x_{IT})$ the hidden vector sequence is specified as $hs = (hs_1, \ldots, hs_{IT})$ and the output vector sequence is given as $ot = (ot_1, \ldots, ot_{IT})$ that are an iteration of *ot* from 1 to *IT*.

$$\vec{hs}_{it} = HS(wt_{xhs} x_{it} + wt_{hshs} hs_{it-1} + c_{hs}) \tag{7}$$

$$ot_{it} = wt_{hsot} hs_{it} + c_{ot} \tag{8}$$

Here, the hidden layer function or an element-wise application of a sigmoid function is termed as *HS*, the weight matrices are denoted as *wt* and *c* indicate the bias vectors. *HS* can be implemented here by adopting logistic sigmoid function as α.

$$ig_{it} = \alpha(wt_{xig} x_{it} + wt_{hsig} hs_{it-1} + wt_{vig} v_{it-1} + c_{ig}) \tag{9}$$

$$fg_{it} = \alpha(wt_{xfg} x_{it} + wt_{hsfg} hs_{it-1} + wt_{vfg} v_{it-1} + c_{fg}) \tag{10}$$

$$v_{it} = fg_{it} v_{it-1} + ig_{it} \tanh(wt_{xv} x_{it} + wt_{hsv} hs_{it-1} + c_v) \tag{11}$$

$$ot_{it} = \alpha(wt_{xop} x_{it} + wt_{hsop} hs_{it-1} + wt_{vop} v_{it} + c_{op}) \tag{12}$$

$$hs_{it} = ot_{it} \tanh(v_{it}) \tag{13}$$

Here, the input gate, forget gate, output gate, and cell activation vectors are represented as ig, fg, op and v, respectively. BRNN [20] includes two different hidden layers that forward to the same output layer. In BRNN, the forward hidden sequence, the backward hidden sequence, and the output sequence are represented as $\vec{hs}$, $\overleftarrow{hs}$ and ot, respectively. The output sequence is updated by iterating $it \in [IT, 1]$ for the backward layer and forward layer iterating $it \in [1, IT]$, which is given in Eqs. (14)–(16).

$$\vec{hs}_{it} = HS(wt_{x\vec{hs}} x_{it} + wt_{\vec{hs}\vec{hs}} \vec{hs}_{it-1} + c_{\vec{hs}}) \tag{14}$$

$$\overleftarrow{hs}_{it} = HS(wt_{x\overleftarrow{hs}} x_{it} + wt_{\overleftarrow{hs}\overleftarrow{hs}} \overleftarrow{hs}_{it-1} + c_{\overleftarrow{hs}}) \tag{15}$$

$$ot_{it} = wt_{\vec{hs}ot} \vec{hs}_{it} + wt_{\overleftarrow{hs}ot} \overleftarrow{hs}_{it} + c_{ot} \tag{16}$$

Here, the higher-level representations can be performed with the developed deep RNN concept by stocking diverse hidden layers on top of each other, with the output sequence of one layer forming the input sequence for the next. For all L layers in the stack, the hidden layer function is employed with the hidden vector sequences hs^l that are iteratively estimated from $l \in [1, L]$ and $it \in [1, IT]$.

$$hs^l{}_{it} = (wt_{hs^{l-1}} hs^l hs_{it}^{l-1} + wt_{hs^l hs^l} hs_{it-1}^l + c_{it}^l) \tag{17}$$

Here, assume $hs^0 = x$ and the network output ot_{it} is computed in Eq. (18).

$$ot_{it} = wt_{hsLot} hs_{it}^L + c_{ot} \tag{18}$$

Each input sequence hs^l is replaced using $\vec{hs}^l$ and $\overleftarrow{hs}^l$ implementing BRNN.

4.3 GOA for Improved RNN

The proposed speech recognition model employs GOA-based RNN for effective classification, where the number of hidden neurons in the RNN classifier is optimized using the GOA. This improves the accuracy and minimizes WER of the proposed speech recognition model. GOA [14] is employed for solving the complexities in the structural optimization of real-time applications. It effectively solves the local optima problems, provides appropriate results, and explores search space. It also improves the accuracy of optimum. It is developed from the swarming behavior of grasshopper insects that are found in both adulthood and nymph. In the nymph phase, the activities of the grasshopper are in small and slow steps, whereas the activities of the adulthood phase are in the extended range and a faster manner. The swarming behavior of the grasshopper is simulated in Eq. (19).

$$Gr_i = So_i + Gf_i + Wa_i \tag{19}$$

In Eq. (19), the position of the i^{th} grasshopper, the social interaction, wind advection, and the gravity force on the i^{th} grasshopper are termed as Gr_i, So_i, Wa_i and Gf_i, respectively. The swarming behavior is modified in Eq. (20) by replacing the values of So_i and Gf_i.

$$Gr_i = \sum_{j=1, j\neq 1}^{N} so(|gr_j - gr_i|) \frac{gr_j - gr_i}{d_{ij}} - C\hat{e}_C + D\hat{e}_w \tag{20}$$

Here, $So_i = \sum_{\substack{j=1 \\ j\neq 1}}^{N} so(|gr_j - gr_i|) \frac{gr_j - gr_i}{d_{ij}}$, where the distance among the i^{th} grasshopper and j^{th} grasshopper, where $d_{ij} = |gr_j - gr_i|$ and the number of grasshoppers is termed as N, and the term $so()$ defines the strength of social forces. $Gf_i = -C\hat{e}_C$, where $\hat{e}_C$ and C denotes a unity vector towards the

center of the earth and the gravitational constant, respectively.$Wa_i = D\hat{e}_w$, in which $\hat{e}_w$ and D represents a unity vector in the direction of the wind and a constant drift, respectively. Eq. (20) is modified for solving the optimization problems that are given in Eq. (21).

$$Gr_i^x = dc\left(\sum_{j=1,j\neq 1}^{N} dc\frac{up_x - lo_x}{2}s(|gr_j^x - gr_i^x|)\frac{gr_j - gr_i}{gr_{ij}}\right) + \hat{T}A_x \tag{21}$$

In Eq. (21), a decreasing coefficient is given as dc that is formulated in Eq. (22), $\hat{T}A_x$ is the value of target at X^{th} dimension, the upper and lower bound in the X^{th} dimension are represented as up_x and lo_x, respectively.

$$dc = dc\max - it\frac{dc\max - dc\min}{IT} \tag{22}$$

In Eq. (22), the maximum value and minimum value of dc are denoted as dcmax and dcmin, respectively, and the maximum number of iterations and the current iteration is noted as IT and it, respectively. The values of dcmax and dcmin are taken as 1 and 0.00001, respectively. The pseudo-code of the GOA algorithm is depicted in Algorithm 1.

Algorithm 1: Grasshopper Optimization Algorithm (GOA)

BEGIN

1. Initialization of grasshopper swarm Gr_i ($i = 1, 2, 3, \ldots, n$);
2. Initialization of variables;
3. Compute the fitness of every search solution agent;
4. Consider the best search solution agent as TA;
5. **While** ($it < IT$) **do**
6. Update dc by Eq. (22);
7. **For each** (search solution agent) **do**
8. Normalization of distances between the grasshoppers;
9. Update the location of the current search agent solution using Eq. (21);
10. **end for**
11. Update if there is the best search solution agent;
12. $it = it + 1$;
13. **end while**
14. Return TA;

END.

Objective function: The proposed speech recognition model on Marathi language using optimized RNN-Based GOA intends to maximize the accuracy and precision for effective recognition. The fitness function of the proposed speech recognition model should maximize accuracy and precision as given in Eq. (23).

$$ff = \arg\max_{\{HN\}}(ACC + \Pr()) \tag{23}$$

Here, the term *ff* denotes the fitness function of the proposed model. Here, the term *HN* number of hidden neurons in the RNN classifier. The solution lies between 5 and 55. The accuracy is represented as *ACC*, and precision is denoted as Pr. Accuracy *ACC* is a "ratio of the observation of exactly predicted to the whole observations" and Precision Pr is "the ratio of positive observations that are predicted exactly to the total number of observations that are positively predicted" that are given in Eqs. (24) and (25), respectively.

$$ACC = \frac{(pot + pon)}{(pot + pon + fap + fan)} \tag{24}$$

$$\Pr = \frac{pot}{pot + pot} \tag{25}$$

Here, the true positives, true negatives, false positives, and false negatives are termed as *pot*, *pon*, *pot*, *fap*, *fan*, respectively.

5 Results and Discussions

5.1 Simulation Setup

The proposed speech recognition model for the Marathi language has been implemented in Python. The proposed model has considered the maximum number of iterations as 25 and the number of populations as 10. It was executed on the Marathi speech corpus that was collected from the "Indian language Technology Proliferation and Development center, Government of India". It has consisted of around 44500 speech files recorded with approximately 1500 speakers with their pronunciation. This speech corpus is collected from all districts of Maharashtra state having different dialects spoken in the state. The performance of the proposed model has been analyzed with the GOA-RNN [21] algorithm by comparing various conventional algorithms such as RNN [20], LSTM [22], and CNN [20]. The Marathi Speech corpus was split into six speech corpora for analyzing the performance.

5.2 Performance Metrics

Various performance measures are used for evaluating the proposed speech recognition model using the optimized RNN with GOA, which is described below.

(a) Sensitivity: "the number of true positives, which are recognized exactly".

$$Sy = \frac{pot}{pot + pon} \tag{26}$$

(b) Specificity: "the number of true negatives, which are determined precisely".

$$Spy = \frac{pon}{fan} \tag{27}$$

(c) FPR: "the ratio of the count of false-positive predictions to the entire count of negative predictions".

$$FPR = \frac{fap}{fap + pon} \tag{28}$$

(d) FNR: "the proportion of positives which yield negative test outcomes with the test".

$$FNR = \frac{fan}{pon + pot} \tag{29}$$

(e) MCC: "correlation coefficient computed by four values".

$$MCC = \frac{pot \times pon - fap \times fan}{\sqrt{(pot + fap)(pot + fan)(pon + fap)(pot + fan)}} \tag{30}$$

(f) WER: It is used for evaluating the proposed speech recognition model.

$$WER(\%) = \frac{De + Ns + Ie}{Nw} \times 100(\%) \tag{31}$$

Here, the terms *Ns*, *De*, *Ie* and *Nw* denotes the number of substitutions in the test, the number of deletions in the test, the number of insertion errors in the test, and the number of words utilized in a test, respectively.

5.3 Performance Analysis

The performance of the conventional and proposed speech recognition model is performed on six speech corpora, and their experiment data is given from Tabs. 2–7. The graphical representation of the performance measure for Speech corpus one is shown in Fig. 2. Also, the GUI for Speech Recognition System allows users to check the model shown in Figs. 3 and 4.

Table 2: Analysis of WER for the proposed speech recognition model for diverse speech corpus

Measures	Speech corpus 1	Speech corpus 2	Speech corpus 3	Speech corpus 4	Speech corpus 5	Speech corpus 6
RNN	12.0276	12.2921	9.6707	12.3456	15.0435	10.2503
LSTM	8.4866	8.6122	6.7623	8.7496	11.971	7.4266
CNN	8.2125	8.2921	6.401	8.3372	11.1171	6.8084
GOA-RNN	7.4177	7.8927	5.8283	7.8564	10.2408	6.5371

Table 3: Overall performance analysis for speech corpus 1

Measures	Accuracy	Sensitivity	Specificity	FPR	FNR	MCC
RNN	0.9369	0.6564	0.9565	0.0434	0.0435	0.5477
LSTM	0.9336	0.6163	0.9563	0.0436	0.3836	0.5211
CNN	0.9425	0.7629	0.9535	0.0464	0.2370	0.5896
GOA-RNN	0.9612	0.7946	0.9681	0.0318	0.2053	0.6158

Table 4: Overall performance analysis for speech corpus 2

Measures	Accuracy	Specificity	Sensitivity	FPR	FNR	MCC
RNN	0.9385	0.7042	0.9533	0.0466	0.3676	0.5548
LSTM	0.9354	0.7042	0.9543	0.0456	0.3436	0.5355
CNN	0.9336	0.6323	0.9550	0.0449	0.3676	0.5271
GOA-RNN	0.9603	0.7823	0.9675	0.0324	0.2176	0.6048

Table 5: Overall performance analysis for speech corpus 3

Measures	Accuracy	Specificity	Sensitivity	FPR	FNR	MCC
RNN	0.9401	0.7005	0.9554	0.0445	0.2994	0.5621
LSTM	0.9368	0.6730	0.9539	0.0460	0.3269	0.5394
CNN	0.9354	0.6346	0.9562	0.0437	0.3653	0.5300
GOA-RNN	0.9610	0.7861	0.0318	0.2138	0.2138	0.6083

Table 6: Overall performance analysis for speech corpus 4

Measures	Accuracy	Specificity	Sensitivity	FPR	FNR	MCC
RNN	0.9368	0.6751	0.9544	0.0455	0.3248	0.5478
LSTM	0.9405	0.6998	0.9561	0.0438	0.3001	0.5654
CNN	0.9397	0.7166	0.9540	0.0459	0.2833	0.5674
GOA-RNN	0.9605	0.7877	0.9675	0.0324	0.2122	0.6073

Table 7: Overall performance analysis for speech corpus 5

Measures	Accuracy	Specificity	Sensitivity	FPR	FNR	MCC
RNN	0.9414	0.7199	0.9552	0.0447	0.2800	0.5707
LSTM	0.9340	0.6335	0.9554	0.0445	0.3664	0.5301
CNN	0.9333	0.6102	0.9564	0.0435	0.3897	0.5178
GOA-RNN	0.9606	0.7771	0.9681	0.0318	0.2228	0.6053

Analysis of WER

The analysis is carried out on the WER measure for showing the efficiency of the proposed speech recognition model, which is given in Tab. 2 for all six-speech corpus. The GOA-RNN is 4.6%, 1.06%, and 0.79%, enhanced than RNN, LSTM, and CNN, respectively. For speech corpus 2, the GOA-RNN is 4.39%, 0.71%, and 0.39% superior to RNN, LSTM, and CNN, respectively. For speech corpus 3, the GOA-RNN is 3.84%, 0.93%, and 0.57% improved than RNN, LSTM, and CNN, respectively. For speech corpus 4, the GOA-RNN is 4.48%, 0.89%, and 0.48% progressed than RNN, LSTM, and CNN, respectively. For speech corpus 5, the GOA-RNN is 4.80%, 1.73%, and 0.87% improved than RNN, LSTM, and CNN, respectively. For speech corpus 6, the GOA-RNN is 3.71%, 0.88%, and 0.57% enhanced than RNN, LSTM, and CNN, respectively. Similarly, the WER for the proposed model is also analyzed for the remaining Speech corpus. Therefore, the proposed speech recognition model with GOA-RNN has established better results based on the WER metric.

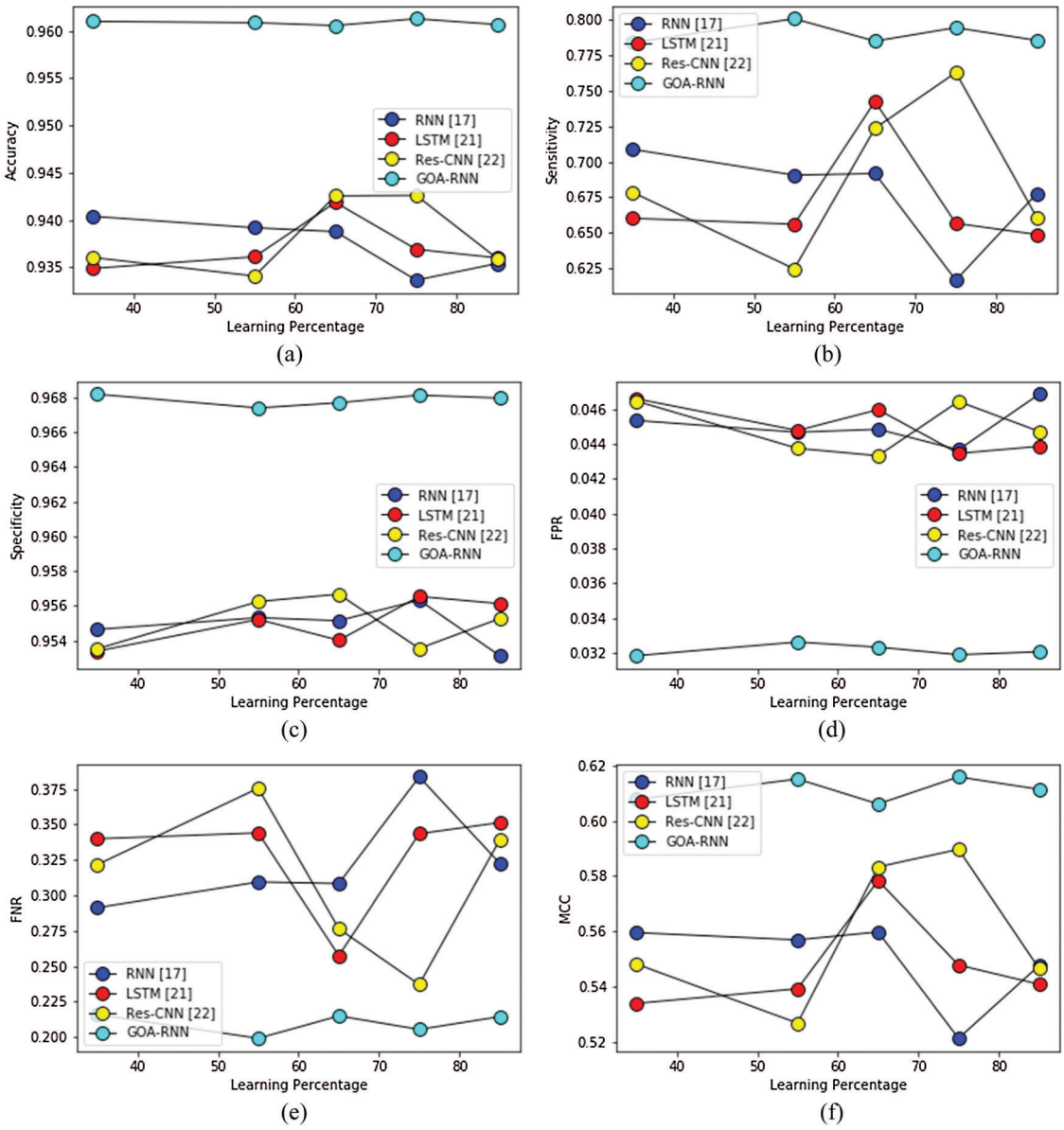

Figure 2: Performance analysis of the proposed and conventional speech recognition model for (a) speech corpus 1 (b) speech corpus 2 (c) speech corpus 3 (d) speech corpus 4 (e) speech corpus 5 and (f) speech corpus 6

Overall Performance Analysis

The proposed speech recognition model is evaluated based on the analysis of accuracy, sensitivity, specificity, FPR, and FNR with six speech corpora, which is given in Fig. 2. As mentioned in the experimental setup, the conventional and proposed algorithms are executed with all six-speech corpus. The evaluation results are given in Tabs. 3–8, and the graphical representation of the results is shown in Fig. 2.

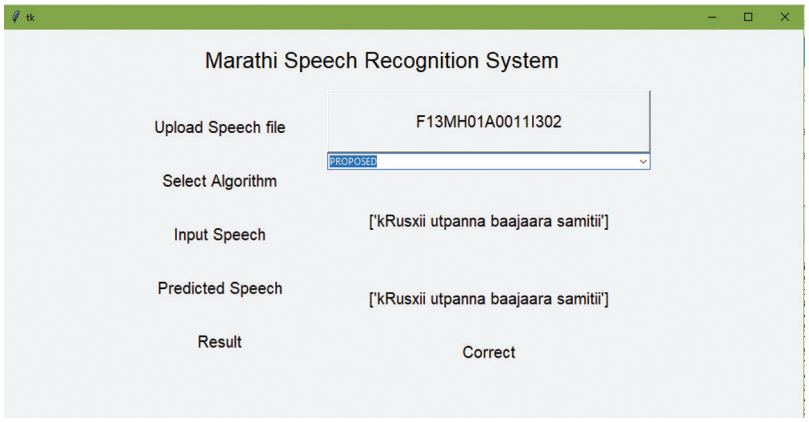

Figure 3: GUI of marathi speech recognition system

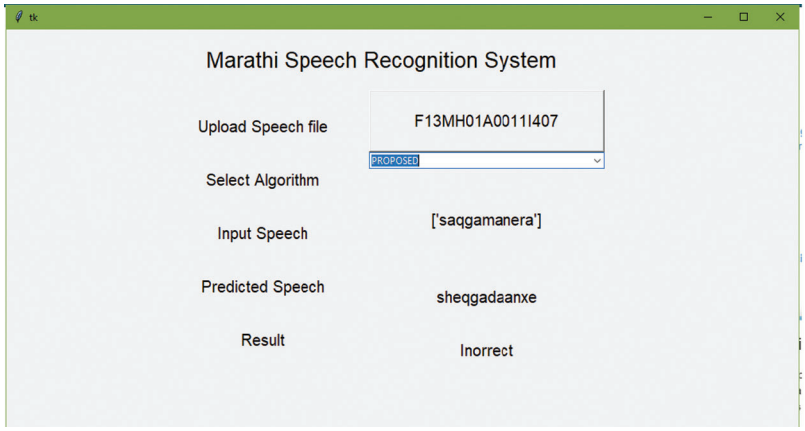

Figure 4: GUI of marathi speech recognition system

Table 8: Overall performance analysis for speech corpus 6

Measures	Accuracy	Specificity	Sensitivity	FPR	FNR	MCC
RNN	0.9370	0.6727	0.9551	0.0448	0.3272	0.5515
LSTM	0.9350	0.6552	0.9542	0.0457	0.3447	0.5360
CNN	0.9390	0.6723	0.9564	0.0435	0.3276	0.5495
GOA-RNN	0.9615	0.8110	0.9675	0.0324	0.1889	0.6180

The results mentioned in Fig. 2. and from Tabs. 3–8 are on a scale of 0 to 1 and explained in percentage while discussing the performance. Therefore, the proposed speech recognition model enhances accuracy when compared to other methods. The parameters used for measuring the performance, such as accuracy, sensitivity, specificity, FPR, and FNR, the proposed optimized RNN algorithm using ROA performed well compared to the RNN, LSTM, and CNN. But the performances of all algorithms for parameter MCC are not satisfactory. Also, the speech recognition system using Python was developed, and the user can validate the different models used here using a graphical user interface shown in Figs. 3 and 4.

Therefore, by considering the overall parameters result, the proposed model has outperformed the other models.

6 Conclusion

This paper has proposed a new speech recognition model for the Marathi language using RNN-based GOA. This model has consisted of three stages such as preprocessing, feature extraction, and classification. The input signals were preprocessed, which was further subjected to the feature extraction stage. Here, the MFCC and spectral-based features were extracted for the proposed speech recognition model. These features were classified using optimized RNN, where the number of hidden neurons was optimized using GOA. Finally, the proposed model has efficiently attained recognized speech. Therefore, from the experimental results, the WER of the proposed model was 3.84%, 1.06%, and 0.79% improved than RNN, LSTM, and CNN, respectively, for speech corpus one, and it has similar results with the remaining speech corpus.

Funding Statement: Taif University Researchers Supporting Project number (TURSP-2020/349), Taif University, Taif, Saudi Arabia.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] 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.

[2] Y. He and X. Dong, "Real time speech recognition algorithm on embedded system based on continuous markov model," *Microprocessors and Microsystems*, vol. 75, pp. 103058, 2020.

[3] T. Aguiar de Lima and M. Da Costa-Abreu, "A survey on automatic speech recognition systems for Portuguese language and its variations," *Computer Speech and Language*, vol. 62, pp. 101055, 2020.

[4] P. Gurunath Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," *Computer Speech and Language*, vol. 63, pp. 101077, 2020.

[5] M. Najafian and M. Russell, "Automatic accent identification as an analytical tool for accent robust automatic speech recognition," *Speech Communication*, vol. 122, no. May, pp. 44–55, 2020.

[6] N. Chatzichrisafis, V. Diakoloukas, V. Digalakis and C. Harizakis, "Gaussian mixture clustering and language adaptation for the development of a new language speech recognition system," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 928–938, 2007.

[7] J. J. Bird, E. Wanner, A. Ekárt and D. R. Faria, "Optimisation of phonetic aware speech recognition through multi-objective evolutionary algorithms," *Expert Systems with Applications*, vol. 153, pp. 113402, 2020.

[8] A. Mohan, R. Rose, S. H. Ghalehjegh and S. Umesh, "Acoustic modelling for speech recognition in Indian languages in an agricultural commodities task domain," *Speech Communication*, vol. 56, no. 1, pp. 167–180, 2014.

[9] A. Becerra, J. I. de la Rosa and E. González, "Speech recognition in a dialog system: From conventional to deep processing: A case study applied to Spanish," *Multimedia Tools and Applications*, vol. 77, no. 12, pp. 15875–15911, 2018.

[10] G. Pironkov, S. U. Wood and S. Dupont, "Hybrid-task learning for robust automatic speech recognition," *Computer Speech and Language*, vol. 64, pp. 101103, 2020.

[11] 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.

[12] J. Guglani and A. N. Mishra, "Automatic speech recognition system with pitch dependent features for Punjabi language on KALDI toolkit," *Applied Acoustics*, vol. 167, pp. 107386, 2020.

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

[14] S. Saremi, S. Mirjalili and A. Lewis, "Grasshopper optimisation algorithm: Theory and application," *Advances in Engineering Software*, vol. 105, pp. 30–47, 2017.

[15] 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.

[16] S. Herzog, "Efficient DSP implementation of median filtering for real-time audio noise reduction," in *DAFx 2013-16th Int. Conf. on Digital Audio Effects*, Maynooth, Ireland, pp. 1–6, 2013.

[17] V. Dogra, A. Singh, S. Verma, N. Z. Jhanjhi and M. N. Talib, "Understanding of data preprocessing for dimensionality reduction using feature selection techniques in text classification," in *Intelligent Computing and Innovation on Data Science. Lecture Notes in Networks and Systems*, Springer: Singapore, vol. 248, pp. 455–464, 2021.

[18] 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.

[19] 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.

[20] A. M. and G. H. Alex Graves, "Speech recognition with deep recurrent neural networks, department of computer science, University of Toronto," in *ICASSP, IEEE Int. Conf. on Acoustics, Speech and Signal Processing-Proc.*, Vancouver, BC, Canada, pp. 6645–6649, 2013.

[21] A. Ann Rufus and L. Kalaivani, "A GOA–RNN controller for a stand-alone photovoltaic/wind energy hybrid-fed pumping system," *Soft Computing*, vol. 23, no. 23, pp. 12255–12276, 2019.

[22] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou *et al.,* "Speech emotion classification using attention-based LSTM," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 27, no. 11, pp. 1675–1685, 2019.

[23] D. N. Le, G. N. Nguyen, H. Garg, Q. T. Huynh, T. N. Bao *et al.,* "Optimizing bidders selection of multi-round procurement problem in software project management using parallel max-min ant system algorithm," *Computers, Materials & Continua*, vol. 66, no. 1, pp. 993–1010, 2021.

[24] W. J. AL-kubaisy, M. Yousif, B. Al-Khateeb, M. Mahmood and D. N. Le, "The red colobuses monkey: A new nature–inspired metaheuristic optimization algorithm," *International Journal of Computational Intelligence Systems*, vol. 14, no. 1, pp. 1108–1118, 2021.

[25] B. Al-Khateeb, K. Ahmed, M. Mahmood and D. Le, "Rock hyraxes swarm optimization: A new nature-inspired metaheuristic optimization algorithm," *Computers, Materials & Continua*, vol. 68, no. 1, pp. 643–654, 2021.