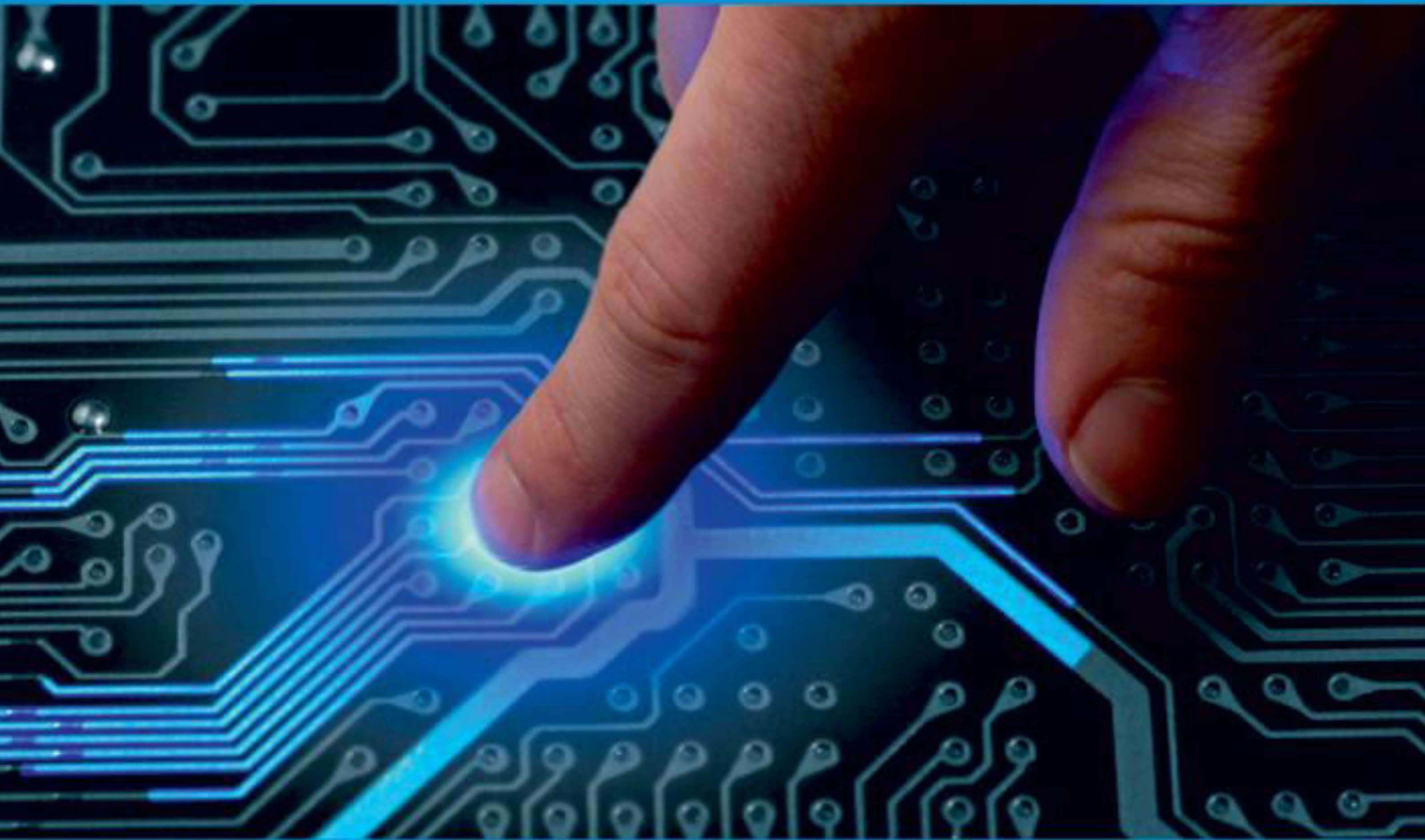
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 7, July 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007084 |

Speech Command Recognition Model based on Adaptive Artificial Bee Colony Optimization and Long Short-Term Memory

Shilpa S[1], Anupama H[2]

Research Scholar, Bangalore Institute of Technology, Bengaluru, Karnataka, India[1]

Assistant Professor Dept. of Electronics & Communication Engineering, AMC Engineering College, Bengaluru, Karnataka, India[2]

ABSTRACT: Speech command recognition is required for the various home appliance for controlling smart devices, text-to-speech conversion and voice command devices. The existing methods in speech command recognition involve applying Generative Adversarial Network (GAN) and Convolution Neural Network (CNN). The GAN and CNN method has a limitation of overfitting problem and this requires unique feature selection to improve its performance. This research applies Adaptive Artificial Bee Colony (AABC) to select features for speech command recognition. The Monte Carlos search process is applied in the ABC method to improve the search efficiency and maintain the exploration-exploitation. Mel Frequency Cepstral Coefficient (MFCC) and Convolutional Neural Network (CNN) feature extractions were applied in the proposed model to extract the features. The Long Short Term Memory (LSTM) model is applied for classification due to its capacity to store sequences of features from MFCC and CNN models. The AABC-LSTM model has the accuracy of 92.4 % and the existing TERA model has 89.7 % accuracy in the speech command dataset.

KEYWORDS: Adaptive Artificial Bee Colony, Convolution Neural Network, Long Short Term Memory, Mel frequency cepstral coefficients, and Speech Command Recognition.

I. INTRODUCTION

Speech communication is considered an efficient mode of communication between a machine and a human. Speech signal translation into text is carried out using speech recognition which is the linguistic field of computational. Methodologies and rule-based technologies are developed to convert spoken acoustic signals into words using computers or mobile [1]. Recently, studies have shown that the location constraint method of attention techniques is required to minimize misrecognition due to incorrect alignments of end to end model with the attention technique of speech recognition systems. Speech recognition systems have the significant advantage of applying location constraint vector to consider monotonicity alignment [2]. The automatic Speech recognition system of stat-of-the-art technique usually separates the speech recognition task into various sub-tasks that optimize independently. Acoustic feature observations such as Perceptual Linear Prediction cepstral features (PLPs) or MFCCs were extracted from speech signals of short term on knowledge of speech perception and speech production [3]. Data augmentation is an effective way to enlarge training data and size to solve this problem. GAN method is applied by some researchers for data generation instead of the traditional method of directly adding noise to the original waveform to improve speech recognition under noise conditions [4].

The speech enhancement method is applied to remove the noise for noise characteristics estimation. The speech enhancement method enhances the input speech signal quality but not speech intelligibility [5]. Some feature extraction techniques use entropy based features, wavelet coefficients, MFCC, and spectrogram etc., [6-7]. Deep learning techniques gained lot of attention recently due to its unparalleled success in various applications such as biomedical engineering and clinical diagnostics. Deep learning methods have the significant advantage of no need to manually craft features from data since the network learns abstract representations of data and useful features from data through training. For instance, the CNN model is generally applied for 2D image signals and Recurrent Neural Networks (RNNs) are more suitable to process time-series signals [8, 9]. Speakers of two groups across formant frequencies and large pitch differences degrade Automatic Speech Recognition tasks performance [10]. The objectives and contribution of this research are discussed as follows:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007084 |

1. The AABC method is proposed to improve the efficiency of the speaker command recognition system. The Monte Carlo technique is applied to increase the search capacity of the ABC method and maintain exploration-exploitation tradeoff.
2. The LSTM model is applied to handle the long-term sequence of MFCC and CNN extracted features. The LSTM model is suitable to handle the sequence of the feature due to its capacity to store the relevant information for the long term.
3. The AABC-LSTM model shows higher efficiency than ABC, existing feature selection, and classifiers. The AABC-LSTM model has higher efficiency than existing methods in the speaker recognition system.

The organization of the paper is given as follows: Literature survey of recent speaker command recognition system is explained in Section 2 and the explanation of the AABC-LSTM model is illustrated in Section 3. The results discussion is given in Section 4. The conclusion of this research work is given in Section 5.

II. LITERATURE SURVEY

Neural Network based speech recognition models were applied to improve the efficiency of speech processing. The recent methods in speech recognition models were reviewed in this section.

Liu, *et al*. [11] developed a Transformer Encoder Representation from Alteration (TERA), a self-supervised speech pre-training model. To train a large amount of unlabelled speech on encoders, three orthogonal axes with alteration were applied in the model. Acoustic frames reconstruction was learned in the model from their altered counterpart and stochastic policy was used to alter along various dimensions: magnitude, frequency, and time. TERA model performs fine-tuning with downstream model or speech representation extraction. The results show that the developed model has higher efficiency than existing methods in speech recognition. The TERA model requires a large amount of data for training the network and has lower efficiency in the small dataset.

Wang, *et al*. [12] applied Conditional Generative Adversarial Network (CGAN) mode to craft adversarial examples of targeted speech. The target label was transformed into a vector and considered as condition input of CGAN. Perturbation was generated using a generator in CGAN to provide an adversarial example of misclassified as a keyword of pre-specified target that the discriminator deceives to misclassify the adversarial example as genuine. Adversarial examples of crafts are differentiated by discriminator from legitimate samples. Ensembled KWS classifiers were applied for the target network and the proposed CGAN framework enforce the generator to develop model-independent perturbation. The developed method has lower efficiency in unconstraint speech length.

Wang, *et al*. [13] applied Generative Adversarial Network (GAN) model for constructing the target speech of adversarial network. The target speech recognition network was integrated intothe GAN network and formulated as a third-part game. GAN generator generates perturbation that makesthe target network misclassified to a specified target and fooling the discriminator. The discriminator was applied to distinguish the crafted network from the genuine samples and the target network classification error was back propagated for generator update using gradient flow. Back propagation was carried out using the target network and gradient update in the generator. The developed method requires a large dataset for training and lower efficiency in a small dataset.

Kamsing, *et al*. [14] proposed a particle filter technique with an adaptive method with gradient descent optimizer to improve the performance of the model. System computational time was reduced based on a pre-trained model for speech recognition classification using the CNN model. The developed method has higher efficiency in speech recognition systems than existing methods. The fine tuning of parameters was carried out to improve the efficiency of the classification. The developed method has lower cross-entropy than the conventional model.

An, *et al*. [15] applied a hybrid network with a delay feedback reservoir based on fusing convolutional or fully connected neural networks in a unified structure of information processing. The convolutional layer was applied for feature extraction in speech signal due to CNN model was more powerful than the fully connected layer. The ResNet residual block similarity with RC's reservoir blocks was used for the validation of data. The developed method has higher efficiency in the speech process than existing methods. The developed method has an overfitting problem in the training of the network that reduces the efficiency.

1 Proposed Method

2 MFCC

The MFCC features provide a sound power spectrum in a short time. Consider speech signal measures FFT and calculate windowed speech signal of squared magnitude [16, 17]. The signal spectrum of pre-emphasizing is used to equalize the unequal perception rate of human hearing at various frequencies. The acquired spectrum of log amplitude mapping is carried out using Mel scale filters integrate with a log power spectrum by applying critical band filters of

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007084 |

the overlapping window. The Cepstral coefficients are measured by considering IDFT and spectral smoothing is executed. The Mel-bin log energies of DCT are carried out using coefficients, as given in equation (1).

$$C_i(t) = \sum_{a=1}^{B} logm_a(t).\cos\left(\frac{i(a-0.5)\pi}{b}\right) \quad (1)$$

Where the number of filter bank functions of a triangular shape is denoted as b. MEL scale is used to change the spectrum, in Mel. The major benefits of MFCC's are coefficients of speech input signal minimize the spectral data into a smaller number of coefficients. Logarithmic function in MFCC model's the human auditory method of loudness perception. This method is easy to compute, simple and relatively fast. This method is the simplest method of auditory processing.

3 CNN Feature extraction

The CNN architecture consists of Convolutional and pooling layers that are connected to a fully connected layer at the end [18 – 21]. A pooling layer of global average is applied in some cases to replace fully connected layer. Regulatory units such as dropout and batch normalization are applied with various mapping functions to optimize the performance of CNN. The CNN components arrangement is important in designing new architecture and to enhance the performance of the model. The role of CNN architecture components is discussed in this section and the CNN architecture model is shown in Figure 1.

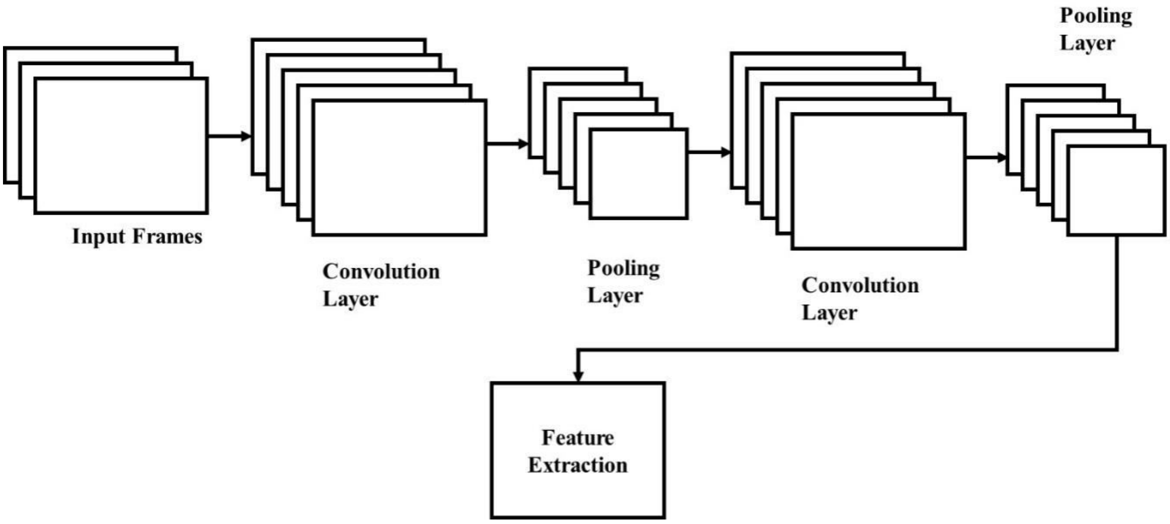


Figure 1. The architecture of CNN model in feature extraction

3.1.1 Convolutional Layer

The kernels set is applied in the convolutional layer where neurons are kernel. If the kernel is symmetric, then convolution operation is a correlation operation. Receptive fields of small slices are divided from the image in a convolutional kernel and an image is divided into small blocks that help to extract motifs features. Kernel uses a specific set of weights to convolve the images by receptive field of corresponding elements which are multiplied by its elements. Equation (2) represents convolution operation.

$$f_l^k(p,q) = \sum_c \sum_{x,y} i_c(x,y).\, e_l^k(u,v) \quad (2)$$

Where k^{th} convolutional operation of feature maps output is expressed as $F_l^k = [f_l^k(1,1), \ldots, f_l^k(p,q), \ldots, f_l^k(P,Q)]$. The k^{th} convolutional kernel k_l of the l^{th} layer of $e_l^k(u,v)$the index is multiplied element-wise and the input image tensor I_c of element is $i_c(x,y)$.

The same set of weights with the sliding kernel is extracted with various sets of features due to the convolutional operation of weight sharing ability. Various types of Convolution operations are convolution direction, padding type and filter set.

3.1.2 Pooling Layer

Convolution operation output is applied to feature motifs and occurs at various parts of the image. The exact location is less important once features are extracted and relative approximate position is preserved. Local operation is based on down-sampling or pooling. Similar information is added in receptive field neighborhodd and local region of dominant response, as in equation (3).

$$Z_l^k = g_p(F_l^k) \quad (3)$$

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165
|| Volume 10, Issue 7, July 2022 ||
| DOI: 10.15680/IJIRCCE.2022.1007084 |

Where type of pooling operation is denoted as $g_p(.)$, k^{th}input feature-map F_l^k in l^{th}a layer of the pooled feature map Z_l^k in pooling operation in equation (2). The pooling operation is applied to extract features combination that is invariant to small distortions and translation shifts. Size reduction of feature map is applied to an invariant feature set that not only regulates network complexity and also increases generalization to reduce overfitting. Various pooling formulations types are L2, max, average, overlapping, and spatial pyramid pooling., which are used in the CNN model.

3.1.3 Activation Function

A decision function is used for the activation function that helps to learn learning patterns of intricate. To accelerate the learning process, an appropriate activation function is used and convolved feature map of the activation function, as given in equation (4).

$$T_l^k = g_a(F_l^k) \tag{4}$$

The convolution output of F_l^k that is applied to the activation function $g_a(.)$ to add non-linearity and obtained a transformed output of T_l^k for l^{th} layer. Various activation functions such as tanh, sigmoid, ReLu were applied for a non-linear combination of features. The ReLu was preferred due to its overcoming of the vanishing gradient problem.

3.1.4 Dropout

Dropout performs regularization in the network to ultimately increase generalization for some units that are randomly skipping with a certain probability. Multiple connections in neural networks learn a non-linear relationship between the features and sometimes this causes overfitting. Network architecture random drop or units helps to select representation with small weights.

3.1.5 Fully Connected Layer

At the end of the classification, the fully connected layer is usually applied. The global operation is applied instead of pooling and convolution. Input is feature extraction preceding layer that are globally analyzed to provide output. A non-linear combination of selected features is applied for data classification.

4 Adaptive Artificial Bee Colony

The foraging behavior of bees is inspired to develop of the ABC method [22, 23]. The ABC algorithm optimizes function extrema value that measures fitness function and is simulated as food sources. The function optimization process of solutions generation is modeled as bees searching for the function of solution space. Bees are three types: scout bees, onlooker bees, and employed bees based on labor divisions.

Food sources are randomly selected by employed bees at the initial stage and neighbors' search is based on greedy selection for selections optimization. Employed bees selected by onlooker bees based on the transformation of feedback information for quality food sources based on neighbour search. Several times, scout bees are selected from employed bees due to food source selection which is not updated, and new scouts search for new food sources. For bee's cooperateto find the best food sources, three types of transformation were used based on optimal extremum value function.

In search space, initial food sources are randomly selected and spread uniformly. Equation (5) denotes the process.

$$x_{ij} = x_j^l + rand(0,1).(x_j^u - x_j^l) \tag{5}$$

Where the total number of the colony is denoted as SN, $i \in \{1,2,\dots,SN\}$, $j \in \{1,2,\dots,n\}$, upper and lower bound of j^{th}dimension is denoted as x_j^u and x_j^l, respectively.

Employed Bees Phase: Initial food sources are searched by each employed bee. This process generates food sources of new solution v_{ij} and update with information of neighbors of its present positionx_{ij}. Equation (6) calculate v_{ij}.

$$v_{ij} = x_{ij} + \varphi_{ij}.(x_{ij} - x_{neighbor}) \tag{6}$$

Where Neighbour $\neq i$, random numbers $j \in \{1,2,\dots,n\}$ and $neighbor \in \{1,2,\dots,SN\}$, and the random number in a range of [-1,1] that evenly distribute control search range. Employed bees are selected as v_{ij} and x_{ij} based on the greedy selection principle. If a new food source is better than the original food sources, means that $fit(v_{ij}) > fit(x_{ij})$, x_{ij} replaced by v_{ij}, x_{ij} is retained.

Onlooker Bees Phase: Employed bees shared food source is selected by each onlooker bee. The fitness value probability is used for selection and fitness value probability is expressed in equation (7).

$$p_i = \frac{fit(x_i)}{\sum_{i=1}^{SN} fit(x_i)} \tag{7}$$

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165
|| Volume 10, Issue 7, July 2022 ||
| DOI: 10.15680/IJIRCCE.2022.1007084 |

Where i[th] food source of fitness value is denoted as $fit(x_i)$. More onlooker bees are searching around food sources of high fitness value and the convergence speed of the algorithm is improved by this process.

Scout Bees Phase: After several iterations, if the food source is not updated, then this source is ignored and the scout bee transformsthe corresponding employed bee. In the global scope, a new food source randomly selects scout bees to continue the neighbor search. New food source selection of generation is described in equation (5).

The Monte Carlo stochastic simulation is used in this research to improve the search performance of the ABC method, as given in equations (8 - 10).

$$x_{id}(t+1) = p_{id}(t) \pm \alpha \times |mbest_d(t) - x_{id}(t)| \times \ln\left(\frac{1}{u}\right) \quad (8)$$

$$p_{id}(t) = \varphi P_{id}(t) + (1-\varphi) G_d(t) \quad (9)$$

$$mbest_d(t) = \frac{1}{M}\sum_{i=1}^{M} P_{id}(t) \quad (10)$$

Where random position between P_{id} and G_d is denoted as p_{id} and a random number between zero and one is denoted as G_d; φ, u. Equation (1) considers the minus sign if $\varphi < 0.5$, otherwise consider this plus sign. Individuals of current optimal position of central point are denoted as $mbest_d$, which is known as mainstream through. The contraction-expansion coefficient is denoted as α. The algorithm convergence rate is controlled by changing α and a higher α is more sufficient for global search and a lower α adapts to the local search at end of the iteration. The control of α is a linear decrease between 1 and 0.4.

5 Long Short Term Memory

LSTM is a special kind of RNN that learns long-term dependencies and remembers information for the long term. A different structure has a repeating module. This model has four interaction layers with unique communication instead of a single neural network [24, 25]. The LSTM unit structure is shown in Figure 2.

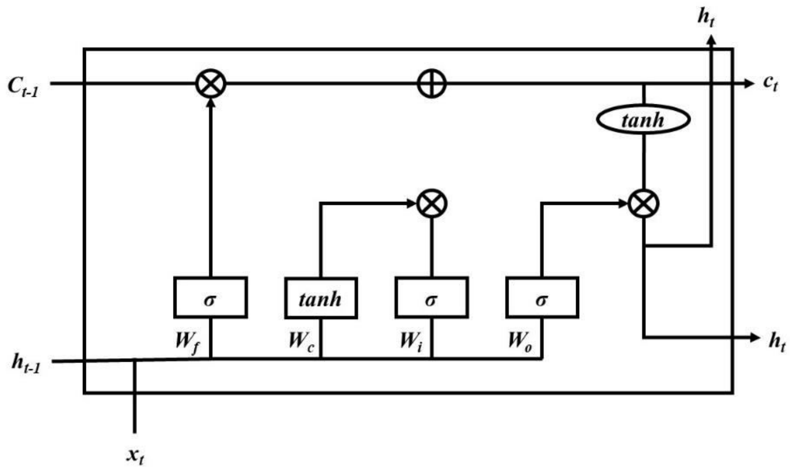


Figure 2. The architecture of the LSTM unit

LSTM network construction is involved in identifying information of irrelevant features and eliminating the information from the cell in that step. The sigmoid function is used to exclude and identify the data that considers the LSTM unit (h_{t-1}) output at time $t-1$ and current input (X_t) at time t. The sigmoid function determines which part needs to be eliminated. Forget gate (or f_t), where f_t is a vector in the range of 0 to 1 related to cell state C_{t-1} of each number, as in equation (11).

$$f_t = \sigma(W_f[h_{t-1}, X_t] + b_f) \quad (11)$$

Where bias and weight matrices are denoted as b_f and W_f, and sigmoid function is denoted as σ.

The new input (X_t) is used to decide and store the information in cell state and cell state update. Two parts are present in this step, the tanh layer, and the sigmoid layer. The new information is needed to update or ignore is determined by the sigmoid layer and tanh function is applied to decide the level of importance (-1 to 1). The two values are multiplied to update the new cell state and new memory results are added tothe old memory C_{t-1} value, as given in equations (12 - 14).

$$i_t = \sigma(W_i[h_{t-1}, X_t] + b_i) \quad (12)$$

$$N_t = \tanh(W_n[h_{t-1}, X_t] + b_n) \quad (13)$$

$$C_t = C_{t-1} f_t + N_t i_t \quad (14)$$

Where cell states at time $t-1$ and t are denoted as C_{t-1} and C_t, the bias and weight matrices are denoted as b and W, respectively.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007084 |

Output cell state (O_t) is applied to update output values (h_t). A sigmoid layer determinesthe output provided by which cell state parts. The sigmoid gate output (O_t) is multiplied by $tanh$ layer new values that are in the range of -1 to 1.

The bias and weight matrices are denoted as b_o and W_o, respectively of the output gate.

6 Results

The AABC-LSTM model is evaluated with quantitative and comparative analysis on speaker recognition.

Table 1. Feature extraction performance analysis

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)
MFCC	81.2	81.5	81.2
CNN	88.4	87.2	87.6
MFCC-CNN	92.4	93.1	93.2

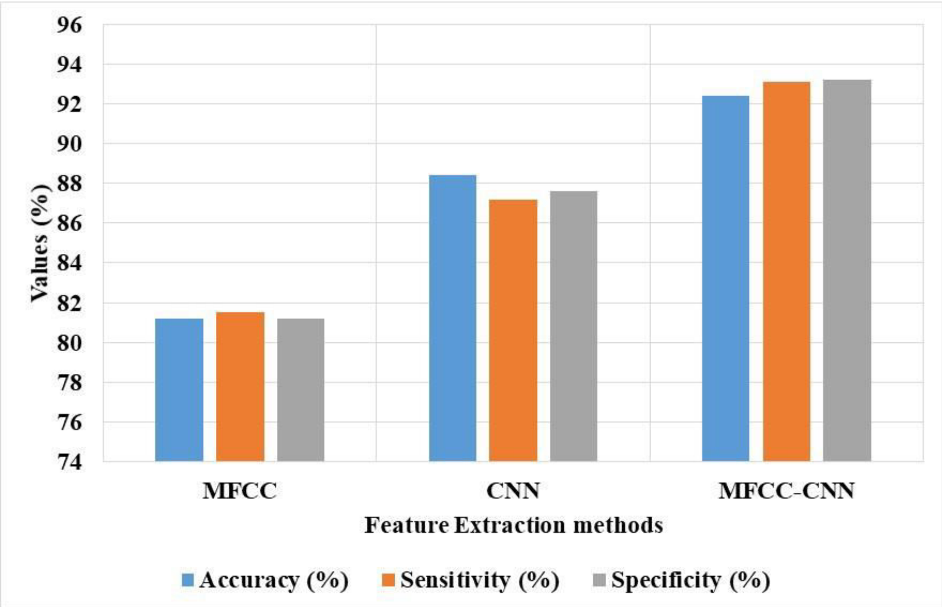


Figure 3. Feature extraction performance analysis on speech recognition

Figure 3 and Table 1 provide the feature extraction method analysis on speaker recognition. This shows that the MFCC features provide considerable performance and CNN increases the performance of speaker recognition. The hybrid feature selection method improves the performance of the speaker recognition due to MFCC method extracts the structural features and CNN extracts the statistical features. The MFCC-CNN model has accuracy of 92.4 %, CNN has 88.4 % accuracy and the MFCC method has 81.2 % accuracy.

Table 2. Feature selection performance analysis

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)
ACO	84.6	83.2	81.6
WOA	83.7	81.4	80.5
ABC	86.3	86.5	86.1
AABC	92.4	93.1	93.2

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165
|| Volume 10, Issue 7, July 2022 ||
| DOI: 10.15680/IJIRCCE.2022.1007084 |

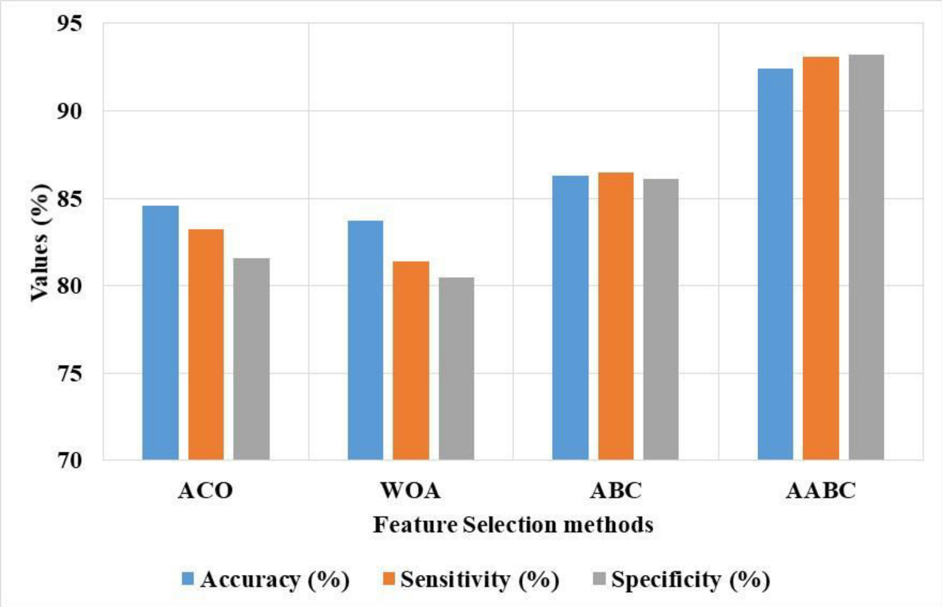


Figure 4. Feature selection method comparison on speaker recognition

The AABC feature selection method is compared with existing feature selection methods in Figure 4 and Table 2. The AABC method has the advantage of applying the Monte Carlo to increase the search efficiency of the model. The AABC method improves the tradeoff between exploration and exploitation of the model. The ABC method has lower exploitation, the WOA method has lower convergence and the ACO model has a local optima trap. The AABC method has accuracy of 92.4 %, ABC has 86.3 % accuracy, and WOA method has 83.7 % accuracy.

Table 3. Classifier performance analysis

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)
SVM	85.3	86.2	86.1
RF	88.7	88.3	86.2
DNN	91.5	91.6	90.4
LSTM	92.4	93.1	93.2

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165
|| Volume 10, Issue 7, July 2022 ||
| DOI: 10.15680/IJIRCCE.2022.1007084 |

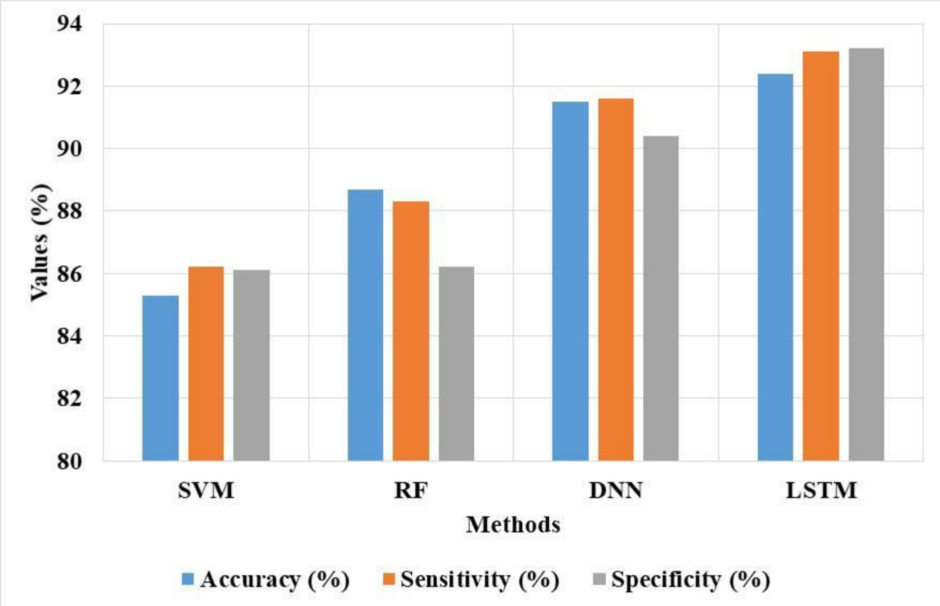


Figure 5. Classifiers' performance analysis on speaker recognition

The AABC-LSTM model is compared with existing classifiers techniques, as given in Figure 5 and Table 3. The LSTM method is suitable to store the sequence of features provided by MFCC feature extraction and effectively handles CNN features. The DNN and Random Forest (RF) model has an overfitting problem in the classification process. The SVM model has a limitation of imbalance data problem that tends to degrade the model performance. The AABC-LSTM model has accuracy of 92.4 %, DNN has 91.5 % accuracy and SVM has 85.3 % accuracy.

Table 4. Comparative analysis on speaker recognition

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)
AABC-LSTM	92.4	93.1	93.2
TERA [11]	89.7	88.7	89.8
CGAN [12]	83.2	82.4	83.5
GAN [13]	85.4	84.7	87.4
particle filter [14]	86.1	85.4	86.7
CNN [15]	88.3	87.3	88.2

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165
|| Volume 10, Issue 7, July 2022 ||
| DOI: 10.15680/IJIRCCE.2022.1007084 |

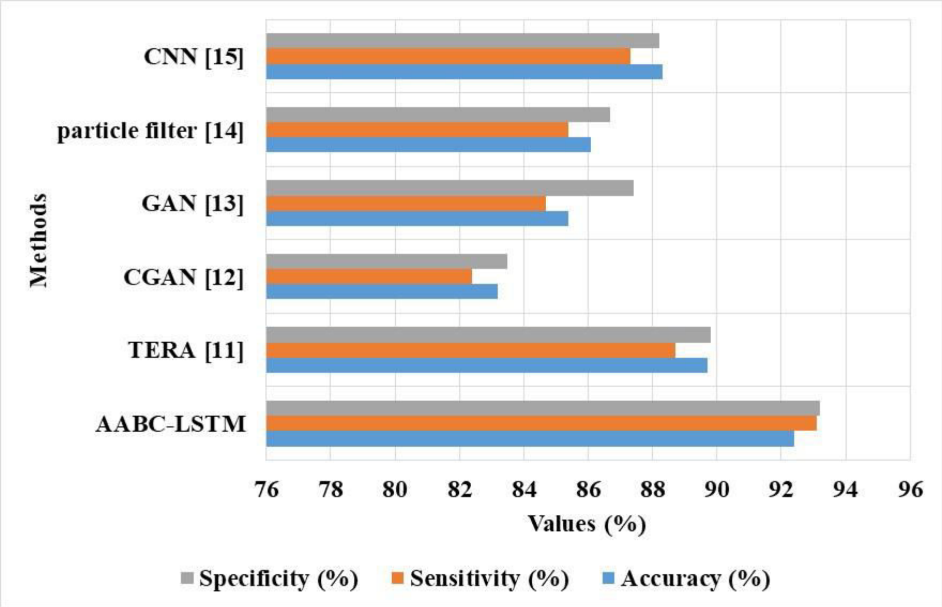


Figure 6. Comparative analysis on speaker recognition

The AABC-LSTM model is compared with existing methods in speaker recognition, as given in Figure 6 and Table 4. The AABC-LSTM model increases the search efficiency based on Monte Carlo method and the LSTM model is suitable to store the sequence of feature information in the network. The GAN-based models [12, 13] have the limitation of lower efficiency in feature handling and generated data is consists of lower relevance. The TERA [11] and CNN [15] models have the limitations of overfitting problem that affects the efficiency of the model. The particle filter [14] technique has the limitation of lower convergence and local optima trap.

IV. CONCLUSION

This research proposes the AABC-LSTM model to improve the efficiency of speaker recognition. Google Speaker Command Recognition dataset was used to evaluate the AABC-LSTM model. The MFCC model extracts structural features and the CNN model extracts statistical features for speaker command recognition. The Monte Carlos method effectively improves the search capacity of the ABC feature selection method and maintains the trade-off between exploration and exploitation. The MFCC and CNN extracted features were applied in the LSTM model for the speaker recognition process. The LSTM model handles the sequence of features and provides efficient classification. The AABC-LSTM model has higher efficiency compared to existing methods in the speaker command recognition. The AABC-LSTM model has accuracy of 92.4 % and the TERA model has 89.7 % accuracy in speaker command recognition. The future work of this research involves applying attention based model to improve the efficiency of classification.

REFERENCES

[1] Bhardwaj, V. and Kukreja, V., 2021. Effect of pitch enhancement in Punjabi children's speech recognition system under disparate acoustic conditions. Applied Acoustics, 177, p.107918.

[2] Xue, J., Zheng, T. and Han, J., 2021. Exploring attention mechanisms based on summary information for end-to-end automatic speech recognition. Neurocomputing, 465, pp.514-524.

[3] Palaz, D., Magimai-Doss, M. and Collobert, R., 2019. End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition. Speech Communication, 108, pp.15-32.

[4] Qian, Y., Hu, H. and Tan, T., 2019. Data augmentation using generative adversarial networks for robust speech recognition. Speech Communication, 114, pp.1-9.

[5] Dua, M., Aggarwal, R.K. and Biswas, M., 2018. Performance evaluation of Hindi speech recognition system using optimized filterbanks. Engineering Science and Technology, an International Journal, 21(3), pp.389-398.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165
|| Volume 10, Issue 7, July 2022 ||
| DOI: 10.15680/IJIRCCE.2022.1007084 |

[6] Acharya, J. and Basu, A., 2020. Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning. IEEE transactions on biomedical circuits and systems, 14(3), pp.535-544.
[7] Wang, X., Chen, X. and Cao, C., 2020. Human emotion recognition by optimally fusing facial expression and speech feature. Signal Processing: Image Communication, 84, p.115831.
[8] An, H., An, Q. and Yi, Y., 2019. Realizing behavior level associative memory learning through three-dimensional memristor-based neuromorphic circuits. IEEE Transactions on Emerging Topics in Computational Intelligence, 5(4), pp.668-678.
[9] Zoughi, T., Homayounpour, M.M. and Deypir, M., 2020. Adaptive windows multiple deep residual networks for speech recognition. Expert Systems with Applications, 139, p.112840.
[10] Yadav, I.C., Shahnawazuddin, S. and Pradhan, G., 2019. Addressing noise and pitch sensitivity of speech recognition system through variational mode decomposition based spectral smoothing. Digital Signal Processing, 86, pp.55-64.
[11] Liu, A.T., Li, S.W. and Lee, H.Y., 2021. Tera: Self-supervised learning of transformer encoder representation for speech. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, pp.2351-2366.
[12] Wang, D., Dong, L., Wang, R. and Yan, D., 2021. Fast speech adversarial example generation for keyword spotting system with conditional GAN. Computer Communications, 179, pp.145-156.
[13] Wang, D., Dong, L., Wang, R., Yan, D. and Wang, J., 2020. Targeted speech adversarial example generation with generative adversarial network. IEEE Access, 8, pp.124503-124513.
[14] Kamsing, P., Torteeka, P., Boonpook, W. and Cao, C., 2020. Deep neural learning adaptive sequential monte carlo for automatic image and speech recognition. Applied Computational Intelligence and Soft Computing, 2020.
[15] An, Q., Bai, K., Liu, L., Shen, F. and Yi, Y., 2020. A unified information perceptron using deep reservoir computing. Computers & Electrical Engineering, 85, p.106705.
[16] Agarwal, G. and Om, H., 2021. Performance of deer hunting optimization based deep learning algorithm for speech emotion recognition. Multimedia Tools and Applications, 80(7), pp.9961-9992.
[17] Siam, A.I., Elazm, A.A., El-Bahnasawy, N.A., El Banby, G.M., El-Samie, A. and Fathi, E., 2021. PPG-based human identification using Mel-frequency cepstral coefficients and neural networks. Multimedia Tools and Applications, 80(17), pp.26001-26019.
[18] Huang, K., Liu, X., Fu, S., Guo, D. and Xu, M., 2019. A lightweight privacy-preserving CNN feature extraction framework for mobile sensing. IEEE Transactions on Dependable and Secure Computing, 18(3), pp.1441-1455.
[19] Zhao, H.H. and Liu, H., 2020. Multiple classifiers fusion and CNN feature extraction for handwritten digits recognition. Granular Computing, 5(3), pp.411-418.
[20] Liu, C., Cheng, G., Chen, X. and Pang, Y., 2018. Planetary gears feature extraction and fault diagnosis method based on VMD and CNN. Sensors, 18(5), p.1523.
[21] Amini, M., Pedram, M., Moradi, A. and Ouchani, M., 2021. Diagnosis of Alzheimer's disease severity with fMRI images using robust multitask feature extraction method and convolutional neural network (CNN). Computational and Mathematical Methods in Medicine, 2021.
[22] Zhang, Y., Rong, Y., Yan, C., Liu, J. and Wu, X., 2018. Kernel estimation of Volterra using an adaptive artificial bee colony optimization and its application to speech signal multi-step prediction. IEEE Access, 7, pp.49048-49058.
[23] Rao, H., Shi, X., Rodrigue, A.K., Feng, J., Xia, Y., Elhoseny, M., Yuan, X. and Gu, L., 2019. Feature selection based on artificial bee colony and gradient boosting decision tree. Applied Soft Computing, 74, pp.634-642.
[24] Sherstinsky, A., 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, p.132306.
[25] Lei, J., Liu, C. and Jiang, D., 2019. Fault diagnosis of wind turbine based on Long Short-term memory networks. Renewable energy, 133, pp.422-432.