ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a1

Al-Noor Journal for Information Technology and Cybersecurity

https://jncs.alnoor.edu.iq/

Integrating Deep Learning and Swarm Intelligence for Speech Recognition: A Review

[1] N M Yousif, [1]O Z Abd Al-majid

[1] Department of Computer Sciences, College of Computer Sciences and Mathematics, University of Mosul

Article information

Article history:
Received February, 15,2025
Revised March, 16,2025
Accepted June 20, 2025

Keywords:
CNN
Deep learning
Swarm Optimization
Optimization
RNN

Correspondence:

Noor Mohammed Yousif
noor.24csp39@student

Abstract

With an emphasis on deep learning and bio-inspired optimization techniques, this paper provides an extensive overview of current developments in voice and emotion detection systems. Advanced recurrent networks like GRU and SVNN, attention-based encoder-decoder frameworks, and hybrid CNN-LSTM architectures are just a few of the models examined in the examined papers. In order to increase robustness, feature extraction methods like MFCC, PLPC, LPCC, and log Mel-filter banks are frequently used in conjunction with data augmentation techniques including speed perturbation, noise injection, and pitch shifting. To enhance feature selection and classifier performance, a number of optimization methods are used, including Particle Swarm Optimization (PSO), Cat Swarm Optimization (CSO), Glowworm Swarm Optimization (GSO), and innovative hybrids like MUPW and GREO. The examined works show state-of-the-art accuracy in a variety of tasks, such as multimodal (audio-visual) recognition systems, Arabic dialect recognition, and emotional speech classification. According to experimental results, there are significant improvements in performance compared to standard models; in certain systems, accuracy rates can approach 99.76%. The increasing efficacy of combining deep learning with intelligent optimization is highlighted in this paper, which also makes recommendations for future developments including transducer-based architectures, real-time adaptation, and domain-specific data augmentation.

DOI: https://doi.org/10.69513/jncs.v2.i1.a1

Introduction

For advanced data analysis, especially in voice recognition and associated optimization tasks, this research mainly uses Swarm Optimization algorithms and Deep Learning approaches. Multi-Layer Perceptrons (MLP) and other Deep Neural Networks (DNNs) are essential for function approximation in the Deep Learning framework because of their deep architecture and regularization methods. In order to overcome the shortcomings of conventional RNNs in capturing long-term dependencies, recurrent neural networks (RNNs), in particular Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs), are used to handle sequential data, such as speech(1-3) . In this study, swarm optimization strategies are essential. Complex issue solutions are optimized using algorithms like Particle Swarm Optimization (PSO). To accomplish optimization objectives, additional techniques are used, such as Golden Ratio Optimization (GRO) and Equilibrium Optimization (EO). To improve search efficiency and convergence, hybrid approaches such as MUPW, which combines PSO with the Whale Optimization Algorithm (WOA), are also investigated (4-6).

When using Deep Learning for Nepali speech recognition, GRUs represent sequential data more well

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a1

than LSTMs, whereas Convolutional Neural Networks (CNNs) extract spatial characteristics. Mel-Frequency Cepstral Coefficients (MFCCs) are utilized for feature extraction, while Connectionist Temporal Classification (CTC) serves as a loss function for training sequence models(2,7, 8).

In addition to Deep Learning and Swarm Optimization, other techniques that aid in feature selection, classification, and hyperparameter optimization include Random Forest, Weighted Binary Cuckoo Search (WBCS), Logistic Regression, Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), and Grey Wolf Optimization (GWO). The procedure includes the study of speech properties including power, rate, formants, and pitch as well as feature extraction methods like Linear Prediction Coefficients (LPC) and Linear Prediction Cepstral Coefficients (LPCC)(9).

Additionally, in addition to techniques like Opposition-Based Learning (OBL), Support Vector Neural Networks (SVNN), SFS-Guided WOA, and the Gravitational Search Algorithm (GSA), other optimization algorithms like Cat Swarm Optimization (CSO) are used for feature selection, improving optimization, classification accuracy, and hyperparameter tuning. Bidirectional Recurrent Neural Networks (BRNNs) and Bidirectional LSTMs (BLSTMs) are used to capture long-range relationships in sequence-based tasks. For certain applications, recurrent neural network language models (RNN-LMs) and artificial neural networks (ANNs) are also pertinent.

1.Deep Learning and Speech Recognition

Alsayadi et al. (2021) (7) investigates advanced deep learning techniques for Arabic automatic speech recognition (ASR), focusing on diacritized speech. For sequence alignment, the authors suggest a CNN-LSTM hybrid model that incorporates CTC (Connectionist Temporal Classification) and an attention-based encoder-decoder. MFCCs and log Mel-filterbank energies are used for feature extraction, and training robustness is improved by data augmentation (speed perturbation, noise injection, pitch shifting). The models are implemented by the Espresso and ESPnet toolkits, which enhance decoding accuracy through shallow fusion by merging external RNN language models (RNN-LM). The CNN-LSTM with attention shown cutting-edge capabilities: WER (word error rate): 28.48% ,5.66% is the Character Error Rate (CER). This fared better than both traditional ASR (WER: 33.72%) and dual CTC-attention (WER: 31.10%).(7).

Alsobhani et al. (2021) (8) proposed a 13-layer CNN model to classify 6 control words recorded in diverse environments. Start, stop, forward, reverse, right, and left are the six control words that are employed. statements made by individuals of various ages.This study suggested turning audio into a spectrogram in the form of a picture. For example, let's say we have an image with dimensions of N = 6 * 6 and the kernel filter is F = 3 * 3. The use of real-world noisy data made the study unique and improved the resilience of the model. The 97.06% accuracy attained shows promise for uses such as voice-activated robotics(10).

Dua et al (2022) (9), proposed a CNN-based model to analyze Punjabi Gurbani hymns with background music.This work also utilizes Praat for speech segmentation and the CNN model, which has six layers of 2DConv, 2DMax Pooling, and 256 dense layer units (Google's TensorFlow service). The MFCC feature extraction technique, which extracts both regular speech and background music features, was used to enforce feature extraction. With 418 recordings from 11 speakers, the study produced a novel dataset with an accuracy of 89.15% and a WER of 10.56%. The model showed promise for assistive technologies and religious applications by outperforming more conventional approaches like DTW and HMM. The following figure1 illustrates the suggested framework (Dua et al. 2022) (9)

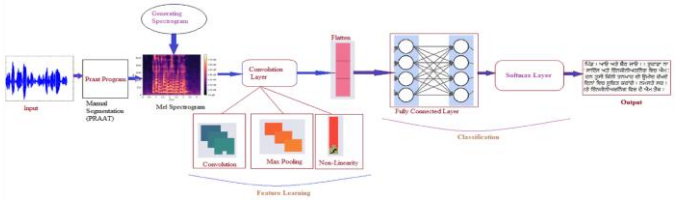

Figure 1. Framework for proposed work.

Shashidhar et al. (2022)(1) presented a novel method called Combining Audio and Visual Speech Recognition Using LSTM and Deep Convolutional Neural Network. An integrated speech recognition system that uses both lip movements and audio signals is presented in this paper. A specific dataset of seven words in English was created for the study, which included 525 video clips from 15 participants.

In audio processing, lip movement sequences were examined using LSTM algorithms, and features were extracted using MFCC. 90% accuracy for audio-only recognition, 71% for visual-only (lip-reading), and 91% for the combined system were attained by fusing the outputs using a Deep Neural Network (DNN). Although single-word recognition difficulties and dataset scalability were among the drawbacks, the results showed that the multimodal system performed better in noisy conditions. Through varied data collection and model development, the authors

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a1

highlighted potential options such sentence-level analysis and improved real-world application(1).

Alsayadi et al. (2022) (7) addressed the difficulties of identifying various Arabic dialects by introducing a hybrid model that combines CNN-LSTM and attention-based encoder-decoder approaches. In addition to feature extraction utilizing MFCC and MFF Bank methodologies, the study included data augmentation techniques such pitch-shifting, noise injection, and speed perturbation. RNN-LSTM was used to create the language model, which was then implemented using the Espresso toolkit.

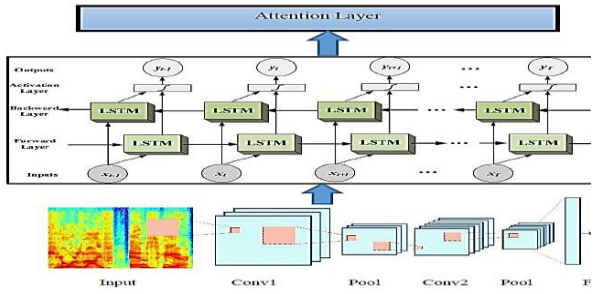

Figure 2. The methodology of employed models.

With a word error rate (WER) of 57.02% and a character error rate (CER) of 25.24%, the model outperformed earlier research and produced encouraging results. The study came to the conclusion that this model is the first of its kind for dialectal Arabic recognition, emphasizing the need of data augmentation in enhancing performance. It also recommended future improvements, such as transducer models and on-the-fly feature extraction(10).

Loan Trinh Van and associates (2022) (11)advocated the use of deep neural networks for emotional speech recognition. In this study, emotions were extracted from audio data in the IEMOCAP database using three deep learning techniques: CNN, CRNN, and GRU. Mel-spectral coefficients and other spectrum and energy-related parameters were used in the study, along with data augmentation methods like voice modulation and white noise addition. With an accuracy of 97.47%, the results showed that the GRU model was better to other models and earlier research. The researchers found that when paired with data augmentation, recurrent neural networks like GRU greatly improve the accuracy of emotion recognition, opening up new possibilities for applications involving human-machine interaction(12).

Rudregowda et al.(2023)(13) an integrated system combining auditory and visual speech processing was developed. The model employed advanced techniques, Lip movements were examined using Dlib and converted to digital form using LSTM networks, while audio features were extracted using Mel-Frequency Cepstral Coefficients (MFCC). A Deep Convolutional Neural Network (DCNN) was utilized to fuse audiovisual information, with 94.67% training accuracy and 91.75% testing accuracy. Findings showed that this hybrid strategy outperformed conventional techniques, opening the door for cutting-edge assistive technology for applications involving human-computer interaction and the hearing impaired(23).

Samuel Thomas et al. (2021) (14) propose an RNN Transducer Models for Spoken Language Understanding. The construction and adaptation of RNN-T models for SLU in three real-world scenarios—the availability of verbatim transcripts, audio data with sparse annotations, or text-only data without speech recordings—were thoroughly examined in this work. To process speech directly into entities and intentions, the study used RNN-T models, which are composed of three sub-networks: transcription, prediction, and joint. Switchboard data was used to pre-train the models on ASR systems, and non-acoustic symbols were added to modify them for SLU tasks. The models' performance on the ATIS and Call Center datasets was close to state-of-the-art, with an entity F1-score of 93.2% and intent recognition accuracy of 94.7%.The outcomes showed that synthetic speech from TTS systems may successfully replace actual recordings, and pre-training is crucial even with sparse audio data. The study found that while initial ASR accuracy has little bearing on ultimate results, SLU performance is mostly dependent on model pre-training(Thomas et al. 2021).(14)

Abdelfatah et al. (2020) (15) developed a system leveraging acoustic analysis of cough, breathing, and voice patterns to distinguish COVID-19 patients from healthy individuals. Six important acoustic features were analyzed by the system using deep learning methods, specifically LSTM networks, with MFCCs performing better. The results demonstrated that respiratory sounds were more reliable than speech for detection, with 98.2% accuracy for breathing sounds and 97% for coughing, while voice analysis obtained 88.2% accuracy. The authors pointed out that, while its effectiveness as a non-invasive screening tool, its limited sample size (80 individuals) and lack of comparative data from other respiratory disorders were drawbacks, and they recommended future extension to improve robustness. The study emphasizes the necessity for diverse datasets while highlighting AI's promise in pandemic response(15).

Wei Zhou et al. (2022) (16) presented a comprehensive framework to enhance the integration of external language models with the internal model in RNN-T systems. In order to improve ILM estimate, the study suggested a unique Exact-ILM Training strategy

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a1

and used Internal Language Model (ILM) correction strategies to resolve the mismatch between these models. Tests on the TED-LIUM and Librispeech datasets showed that the h′mini-LSTM approach performed better with precise training, with word error rates (WER) of 13.2% and 1.8%, respectively. The results of the investigation showed that increasing label probabilities against blanks and rebalancing label distributions lead to improved performance. While highlighting the necessity of careful parameter adjustment to maximize the balance between important impacting elements, the study came to the conclusion that exact ILM correction is crucial for successful language model integration, especially in cross-domain activities. The results provide scalable solutions for both in-domain and cross-domain speech recognition systems, bridging theoretical developments with real-world applications(16).

Ragheb et al. (2021) (3) proposed a recurrent neural network models (RNN, BRNN, LSTM, BLSTM) for speech syllable classification using the TIMIT database. Using MFCC features with derivatives and vector quantization, the study found that a 5-layer BRNN model (30,30,20,25,25 nodes) outperformed conventional HMM models (81.01%) with a peak accuracy of 92.6%. With BLSTM, vowels had the best classification accuracy (98.5%), but plosives were the hardest to classify (66% accuracy). The study revealed the drawbacks of vector quantization in comparison to full MFCC features in maintaining acoustic discriminability, but it also showed the advantages of adding hidden layers and the superiority of bidirectional RNN architectures. The best model selection for fine-grained phonetic classification tasks in ASR systems is greatly advanced by these findings(3).

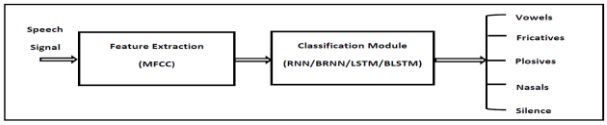

Figure 3. Block diagram of the proposed model.

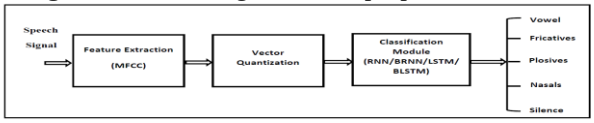

Figure 4. Block diagram of the proposed model with VQ.

Alsayadi (2021) (7) proposed a explored data augmentation for Arabic speech recognition using end-to-end deep learning. The study used an attention-based model for the decoding stage and a hybrid model that used Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to create the acoustic model. To increase the size of the training dataset and improve system performance, data augmentation methods such pitch-shifting, speed transformation, and noise injection were used. To further hone the findings, language models such as RNN-LM and LSTM-LM were employed. The results showed that the suggested model was more effective than conventional techniques, with a notable decrease in Word Error Rate (WER) of 4.55% and Character Error Rate (CER) of 0.96%.The study came to the conclusion that data augmentation is essential for raising the accuracy of Arabic voice recognition systems, and that more improvements may be made by creating increasingly complex data augmentation methods and integrating them with cutting-edge models(7).

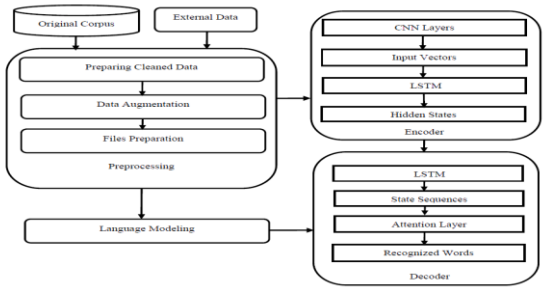

Figure 5. The proposed System Architecture.

Oruh et al. (2022) (17) developed an improved LSTM-RNN model to tackle voice recognition issues by using RNN as a forget gate to reset cell states in sub-sequences, increasing the processing efficiency of continuous input streams. The researchers retrieved MFCC features from the spoken English digit dataset and used the Adam optimization algorithm to fine-tune the parameters. With the loss decreased to 0.02656 at a modest learning rate, the model outperformed other models such as ResNet and DenseNet, achieving an astounding accuracy of 99.36%. The findings showed that LSTM-RNN is still the best option for voice recognition tasks because of its exceptional capacity to represent both short-term and long-term dependencies, with the possibility of increasing accuracy by adjusting learning parameters(17).

Vorontsova et al. (2021) (2) proposed a developed a model for silent speech recognition using electroencephalography (EEG) signals. The study used recordings of 270 healthy participants silently articulating nine words—eight Russian words and one pseudoword—and sophisticated data processing methods, including Fast Fourier Transform (FFT), Independent Component Analysis (ICA) for eye noise filtering, and dimensionality reduction using PCA and t-SNE. In order to analyze spatiotemporal data, a hybrid approach that combines ResNet18 with GRU recurrent units was created. This algorithm achieved 88% accuracy in binary classification and 85% accuracy in 9-word classification. The findings cleared the path for the development of brain-computer

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a1

interfaces (BCIs) for individuals with disabilities by showing that individual data produced more accurate results than group data and confirming the presence of consistent brain activity patterns during silent speaking(2).

1. Swarm Optimization and Speech Recognition Fatemeh Daneshfar et al.(2019) (18) proposed an enhanced Speech Emotion Recognition (SER) system Using methods such as MFCC, PLPC, PMVDR, and their first and second derivatives, a rich high-dimensional feature vector was recovered from glottal and speech waveforms. pQPSO is a modified version of the Quantum-behaved Particle Swarm Optimization (QPSO) algorithm that is used to optimize the Gaussian Mixture Model (GMM) classifier's parameters and reduce dimensionality. The pQPSO algorithm outperformed both contemporary approaches like deep neural networks (DNNs) and traditional dimensionality reduction techniques like PCA, LDA, and FA in the evaluation of the suggested system utilizing the EMO-DB, SAVEE, and IEMOCAP databases Its accuracy rates on IEMOCAP were 74.80%, SAVEE was 60.79%, and EMO-DB was 82.82%. Accuracy of recognition was greatly improved by combining glottal and speech data with an effective dimensionality reduction technique. The model provides performance that outperforms several state-of-the-art methods and is nevertheless appropriate for real-time applications, despite a comparatively lengthy training period(18).

Rajasekhar et al. (2019) (19) proposed a two-stage system for emotion recognition: Using features like Non-Negative Matrix Factorization (NMF) and pitch, which were reduced via Principal Component Analysis (PCA) and classified using an Adaptive Fuzzy Classifier (FC), emotion recognition was carried out after gender was first determined using pitch features classified by the k-Nearest Neighbors (k-NN) algorithm. The Glowworm Swarm Optimization (GSO) algorithm was used to optimize the fuzzy classifier's membership function limits. The RAVDESS database, which encompasses eight emotional categories, was used to assess the model. Grey Wolf Optimization (GWO), Firefly (FF), Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC), and Genetic Algorithm (GA) were among the popular algorithms that were surpassed by the suggested GSO-FC model, which achieved a noteworthy accuracy of 91.01%. It was also quite good at other performance parameters like false discovery rate, negative predictive value, sensitivity, specificity, and precision. Dual features (NMF and pitch), dimensionality reduction, and GSO-optimized fuzzy classification, according to the study's findings, greatly improve emotion recognition. The model exhibits high accuracy and robustness, making it a solid contender for intelligent interactive systems(19).

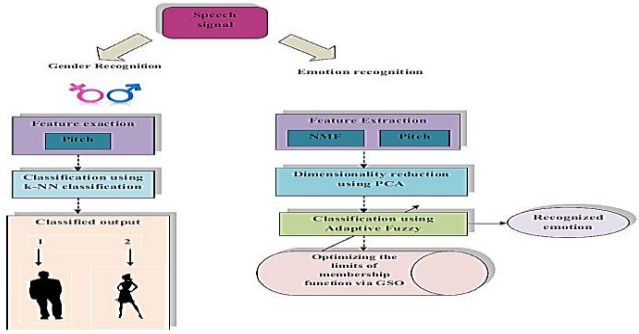

Figure 6. Design of proposed speech emotion recognition model.

Gomathy. (2020) (20) proposed a novel framework for optimal feature selection in speech emotion recognition, utilizing a multi-step process that starts with feature extraction using acoustic features like power, rate, formants, and pitch as well as techniques like MFCC, LPC, and LPCC. The next step was feature selection using Cat Swarm Optimization (CSO) supplemented with Opposition-Based Learning (OBL), which reduced redundancy in the chosen features and increased search efficiency. The framework used a Support Vector Neural Network (SVNN) to classify emotions. When compared to CSO-SVNN and PSO-SVNN, the suggested model performed better on all important metrics: sensitivity of 74% vs 69% and 70%, specificity of 97% versus 93% and 92%, accuracy of 96% versus 89% and 91%, and recognition rate of 93.4% versus 84.3% and 88.8%. By carefully choosing the most pertinent features, the study found that using Enhanced CSO with OBL greatly increases detection accuracy and lowers computing cost, making the model a promising option for implementation in interactive emotion recognition systems(20).

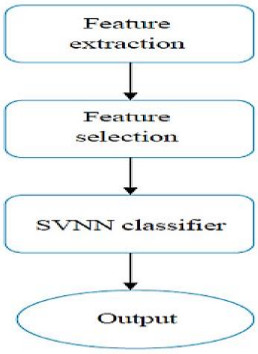

Figure 7. Overview of the proposed approach

Amarjeet Singh, (2020) proposed employing a holistic approach that begins with feature extraction using techniques like MFCC, LPC, and LPCC, together with additional characteristics like speech rate, frequency, and energy, to create a system for accurate speech

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a1

emotion recognition. The Cat Swarm Optimization (CSO) method was used to guarantee the best feature selection, and Opposition-Based Learning (OBL) was added to increase search effectiveness and get rid of redundant features. A Support Vector Neural Network (SVNN) was used to classify emotions. With an accuracy of 96% compared to 89% and 91%, sensitivity of 74% versus 69% and 70%, and specificity of 97% compared to 93% and 92%, the model, which was created in MATLAB, outperformed CSO-SVNN and PSO-SVNN.According to the study's findings, the improved CSO algorithm with OBL considerably increases recognition accuracy, decreases feature redundancy, and lowers computing cost, making the suggested system a more practical and efficient option than conventional techniques(Singh 2020).

Yuan et al. (2021) (22) proposed a speech recognition system for railway dispatch training simulations utilizing a Hidden Markov Model (HMM), improved with the use of a method called Adaptive Escape Particle Swarm Optimization (AEPSO). In order to improve exploration capabilities and prevent local optima during model training, AEPSO dynamically modifies particle velocity and escape tendencies by imitating biological characteristics. The HTK toolkit was used to construct the system, and five speakers—four male and one female—provided real-world voice data. With a word correctness rate of 95.04% versus 94.11%, word accuracy of 91.50% versus 89.26%, and keyword accuracy of 95.13% versus 92.21%, the improved AEPSO-Baum-Welch (AEPSO-BW) model outperformed the conventional Baum-Welch algorithm. Additionally, there was a little rise in deletion error but a decrease in substitution and insertion error rates. In railway dispatcher training simulations, the study found that combining HMM with AEPSO greatly increases voice recognition accuracy, enabling more trustworthy autonomous evaluation and improved system efficiency(23, 24).

Zhang. (2020) (24) proposed a hybrid two-stage system for speech emotion recognition. In the initial phase, each voice feature's significance was assessed using the Random Forest method. Based on calculated weights, the ideal subset of features was chosen in the second step using the Weighted Binary Cuckoo Search (WBCS) method. Several classification algorithms, including Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), and Logistic Regression, were used to the EmoDB dataset. The F1-score metric was used to assess the model's performance. With an F1-score of 0.9822, the data demonstrated that the combination of WBCS and logistic regression performed the best, surpassing all other conventional techniques examined. According to the study's findings, the weighted binary cuckoo method combined with feature importance evaluation increases model efficiency by lowering the amount of features required and improving emotion identification accuracy. The researcher suggested using such methods in other domains like healthcare, education, and safe driving(24).

JYuan et al. (2022) (22)proposed a developed an intelligent speech recognition system to detect lightning whistler waves by utilizing a Long Short-Term Memory (LSTM) neural network, with automatic hyper parameter tuning achieved through the Grey Wolf Optimization (GWO) algorithm. In order to extract auditory elements from the wave audio signals, Mel-Frequency Cepstral Coefficients (MFCC) were utilized. Specifically, the learning rate and the number of hidden units—two crucial LSTM hyperparameters—were optimized using the GWO approach. Three primary steps comprised the recognition process: LSTM training with the optimal parameters, hyperparameter tuning by GWO, and model evaluation on a specified test set. With a test accuracy of 95.5% as opposed to the initial 93.5%, the suggested GWO + LSTM model outperformed the baseline LSTM model by a significant margin. It also showed gains of about 2% in other performance metrics like F1-score, AUC, precision, and recall.The study found that using GWO for hyperparameter tuning greatly increases the model's accuracy and efficiency, making the GWO + LSTM framework ideal for real-time space weather monitoring applications and satellite-based lightning whistler wave identification(22).

Li Zhang et al. (2022) (24) proposed an intelligent sound classification system built on a Convolutional Bidirectional LSTM (CBiLSTM) network, whose hyperparameters were optimized using an enhanced Particle Swarm Optimization (PSO) algorithm. The Short-Time Fourier Transform (STFT) is used to generate spectrograms first, and then a PSO algorithm enhanced with Newton-Raphson and Secant methods, super-ellipse-based hybrid leader generation, and 3D spherical search coefficients is used to tune hyperparameters, such as the number of filters, BiLSTM units, learning rate, and weight decay. Then, using majority voting, three independently optimized CBiLSTM models were combined in an ensemble manner. The suggested system performed noticeably better than default and manually adjusted deep models as well as a number of cutting-edge optimization algorithms when tested on several difficult datasets, including ICBHI 2017 for respiratory sounds, PhysioNet/CinC 2016 for heart sounds, and ESC-10 for environmental sounds. Additionally, in a variety of optimization circumstances, the enhanced PSO variant showed statistical superiority.According to the study's findings, combining CBiLSTM networks with the suggested PSO enhancements allows for strong

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a1

performance on a variety of audio classification tasks, robust optimization without stagnation, and efficient extraction of spatial-temporal features, making it highly applicable in both the medical and environmental domains(24).

3.Integrating Deep Learning and Swarm ptimization for Speech Recognition

Dey et al. (2020) (6)proposes a novel hybrid feature selection algorithm called GREO, It blends Golden Ratio Optimization (GRO) and Equilibrium Optimization (EO), two meta-heuristic optimization techniques. The work detects emotions using the XGBoost classifier and extracts speech components using LPC and LPCC. The hybrid strategy incorporates AWCM (Average Weighted Combination Mean) and SOPF (Sequential One Point Flipping) to further optimize the feature selection. The model achieved classification accuracies of 97.31% and 98.46%, respectively, when tested on two benchmark datasets, SAVEE and EmoDB. The results show that by successfully lowering feature dimensionality while enhancing classification performance, the suggested GREO algorithm beats a number of well-known feature selection methods, making it a promising option for voice emotion identification and other related applications(6).

Rajasekhar B et al.(2020) (19) proposes a presents an innovative spebech emotion recognition (SER) framework that integrates both gender and emotion recognition. The model uses Principal Component Analysis (PCA) to reduce dimensionality after extracting data like pitch and non-negative matrix factorization (NMF). A hybrid approach called MUPW, which combines the Whale Optimization approach (WOA) and Particle Swarm Optimization (PSO), is used to improve the weights of a Deep Belief Network (DBN) for classification. According to experimental results, MUPW performs noticeably better in terms of accuracy, sensitivity, and error reduction than conventional techniques like PSO, GA, WOA, and Firefly. With a maximum accuracy of 97.09%, the model's exceptional performance across a number of metrics was validated by statistical analysis. The authors draw the conclusion that the efficacy of SER systems, especially in applications involving human–computer interaction, can be significantly increased by combining classification models with sophisticated optimization approaches like MUPW(B, M, and V 2020)

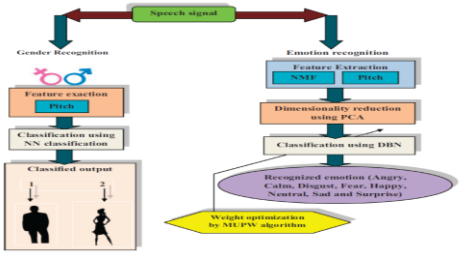

Figure 8. Art of proposed Speech recognition model.

Susmitha Vekkot et al.(2020) (5) presents a hybrid framework for emotional voice conversion while preserving speaker identity. When there is a lack of emotional training data for a given target speaker, this model is especially helpful. The system uses a number of methods, such as intensity modification, wavelet synchrosqueezed transform (WSST) for fundamental frequency (F0) analysis, and mel-generalized cepstral (MGCEP) feature extraction. A particle-swarm-optimized (PSO) artificial neural network (ANN) is used to simulate F0, and a deep neural network (DNN) with speaker adaption is used to map spectral characteristics. Using three multilingual datasets (EmoDB, IITKGP, and SAVEE), the framework demonstrated outstanding performance, with an average mel-cepstral distortion (MCD) of 4.98, F0 RMSE

of 10.67, CMOS score of 3.57, and speaker similarity score of 3.70.According to the study's findings, the hybrid model offers great promise for multilingual speech-to-speech systems and applications involving human-computer interaction by dramatically enhancing emotional voice conversion performance while maintaining speaker characteristics(Vekkot et al. 2020). (5).

Rajeshwar (2021) (25) proposed an innovative model for Speech Emotion Recognition (SER) based on a Deep Belief Network (DBN), with weight optimization by the use of a hybrid method that blends the Gravitational Search method (GSA) with Particle Swarm Optimization (PSO). By extracting pitch and Non-Negative Matrix Factorization (NMF) features, using Principal Component Analysis (PCA) to reduce dimensionality, then classifying using DBN, the model seeks to identify both gender and emotion from speech signals. The findings showed that the suggested strategy achieves up to 20% greater accuracy than conventional techniques like PSO, GA, WOA, and FF. According to the study's findings, the proposed model is particularly appropriate for applications involving human-computer interaction because it is accurate and efficient at automatically recognizing emotions and gender(25).

Alsayadi et al. (2022) (7) proposed an advanced model for speech emotion recognition using a hybrid deep neural network combining Convolutional Neural

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a1

Network (CNN) and Long Short-Term Memory (LSTM), using a Stochastic Fractal Search-guided Whale Optimization Algorithm (SFS-guided WOA) for hyperparameter optimization. In order to improve the dataset, the model uses log-Mel spectrogram characteristics as input and applies a data augmentation approach that generates controlled noise. Using sophisticated metaheuristics, important hyperparameters like learning rate and label smoothing regularization factor are precisely adjusted to improve performance. The experimental findings outperformed all baseline methods with amazing accuracy: 99.47% on RAVDESS, 99.76% on Emo-DB, 99.50% on SAVEE, and 98.13% on IEMOCAP. The study comes to the conclusion that a reliable and extremely accurate system for automatic speech emotion recognition can be achieved by combining deep learning architectures with parameter adjustment and data augmentation (7).

Table 1 Speech recognition techniques

#	Authors (Year)	Focus	Techniques / Models	Dataset	Results	Remarks
1	Daneshfar et al. (2019)	Emotion recognition	MFCC + PLPC + PMVDR + pQPSO + GMM	EMO-DB, SAVEE, IEMOCAP	82.8%, 60.8%, 74.8%	High accuracy with glottal features
2	Rajasekhar et al. (2019)	Emotion recognition with gender	k-NN + NMF + PCA + GSO-FC	RAVDESS	Accuracy: 91.01%	Outperforms PSO, GA, etc.
3	Gomathy (2020)	Feature selection in SER	MFCC + CSO-OBL + SVNN	Not specified	Accuracy: 96%	Better than PSO-SVNN
4	Singh (2020)	SER	MFCC + CSO-OBL + SVNN	Not specified	Accuracy: 96%	Similar to Gomathy
5	Dey et al. (2020)	Emotion recognition	GREO (EO + GRO) + XGBoost	SAVEE, EmoDB	97.31%, 98.46%	Strong feature selection
6	Rajasekhar B et al. (2020)	Gender & emotion recognition	NMF + PCA + DBN + MUPW	Not specified	Accuracy: 97.09%	MUPW outperforms metaheuristics
7	Vekkot et al. (2020)	Voice conversion	MGCEP + WSST + DNN + PSO-ANN	EmoDB, IITKGP, SAVEE	MCD: 4.98, CMOS: 3.57	Preserves speaker identity
8	Hassan et al. (2020)	COVID-19 detection	LSTM + MFCC	80 participants	98.2% (breath), 97% (cough)	Effective screening tool
9	Zhang, Z. (2020)	SER	Random Forest + WBCS	EmoDB	F1-score: 0.9822	Strong hybrid selection
10	H. A. Alsayadi et al. 2021	Arabic ASR (diacritized)	CNN-LSTM + Attention + CTC	Not specified	WER: 28.48%, CER: 5.66%	Outperforms conventional ASR
11	Alsobhani et al. (2021)	Command word recognition	13-layer CNN + Spectrogram	Noisy environments	Accuracy: 97.06%	Robust performance
12	Thomas et al. (2021)	Spoken Language Understanding	RNN-T + Pretraining	ATIS, Call Center	Intent: 94.7%, F1: 91.2%	Pretraining essential
13	Ragheb et al. (2021)	Syllable classification	RNN, LSTM, BLSTM + MFCC	TIMIT	BLSTM: 98.5%	Vowels best, plosives hardest
14	H. Alsayadi et al. 2021	Arabic ASR + augmentation	CNN-LSTM + Attention + Aug.	Not specified	WER ↓ 4.55%, CER ↓ 0.96%	Augmentation helps
15	Vorontsova et al. 2021	Silent speech via EEG	ResNet18 + GRU + ICA/FFT/PCA	270 participants	85% (9-word), 88% binary	Promising for BCIs
16	Liangpan et al. (2021)	Railway ASR	HMM + AEPSO	5 speakers	Accuracy: 91.50%	AEPSO improves over BW
17	Rajeshwar (2021)	Emotion & gender recognition	NMF + PCA + DBN + PSO-GSA	Not specified	Accuracy ↑ 20%	Highly effective hybrid
18	Dua et al. (2022)	Punjabi hymn analysis	CNN + MFCC + TensorFlow	418 recordings	Accuracy: 89.15%, WER: 10.56%	Beats HMM, DTW
19	Shashidhar et al. (2022)	Audio-visual ASR	LSTM + DNN fusion	525 clips	Audio: 90%, Combined: 91%	Good in noisy settings
20	H. A. Alsayadi et al. 2022	Dialectal Arabic ASR	CNN-LSTM + Attention + RNN-LM	Not specified	WER: 57.02%, CER: 25.24%	First for dialectal Arabic
21	Trinh Van et al. (2022)	SER	CNN, CRNN, GRU + MFCC	IEMOCAP	GRU: 97.47%	GRU best performer
22	Zhou et al. (2022)	RNN-T LM integration	h'mini-LSTM + Exact ILM	LibriSpeech, TED-LIUM	WER: 1.8%, 13.2%	Improved LM fusion
23	Orhi et al. (2022)	Digit recognition	LSTM-RNN + MFCC	Spoken digit set	Accuracy: 99.36%	Beats ResNet, DenseNet
24	Yuan et al. (2022)	Lightning wave detection	LSTM + GWO	Custom audio	Accuracy: 95.5%	Beats baseline LSTM
25	Zhang, L. et al. (2022)	Sound classification	CBiLSTM + Enhanced PSO	ICBHI, ESC-10, PhysioNet	-	Strong in medical use
26	Abdelhamid et al. (2022)	SER	CNN-LSTM + SFS-guided WOA	RAVDESS, EmoDB, SAVEE	Up to 99.76%	Best performance overall
27	Rudregowda et al. (2023)	Audio-visual hybrid ASR	MFCC + LSTM + DCNN	Custom dataset	Train: 94.67%, Test: 91.75%	Assistive tech focus

Conclusion

In particular, for languages like Arabic and underrepresented dialects, this work highlights the increasing importance of deep learning and intelligent optimization strategies in improving automatic speech recognition (ASR) and speech emotion recognition (SER) systems. Significant accuracy gains have been shown when hybrid models, including CNN-LSTM with attention mechanisms and encoder-decoder architectures, are included. This is particularly true when paired with external RNN-based language models and Connectionist Temporal Classification (CTC). Furthermore, it has been demonstrated that the regular use of data augmentation techniques like pitch shifting, speed perturbation, and noise injection improves the models' resilience and generalization capacities under a variety of speech situations. According to the reviewed literature, bio-inspired optimization techniques like Whale Optimization Algorithms (WOA), Cat Swarm Optimization (CSO), and Particle Swarm Optimization (PSO) are essential for increasing model efficiency because they minimize computational complexity, optimize hyperparameters, and choose the most pertinent features. When compared to more conventional approaches, these strategies have continuously produced better results. In the end, this corpus of work demonstrates the viability and efficiency of fusing deep learning architectures with intelligent optimization in both scholarly and practical applications, ranging from healthcare and assistive technologies to multilingual speech recognition and emotion analysis. Future studies are urged to investigate on-the-fly feature extraction, multilingual and multimodal systems, and lightweight models that may be implemented in embedded and mobile systems.

Reference

1.Shashidhar, R., S. Patilkulkarni, and S. B. Puneeth. 2022. "Combining Audio and Visual Speech Recognition Using LSTM and Deep Convolutional Neural Network." *International Journal of Information Technology (Singapore)* 14 (7): 3425–36. https://doi.org/10.1007/s41870-022-00907-y.

2.Vorontsova D, Ivan M, Aleksandr Z, Kirill O, Peter R, Ekaterina Zvereva, Lev F. Silent Eeg-Speech Recognition Using Convolutional and Recurrent Neural Network with 85% Accuracy of 9 Words Classification." *Sensors* 2021;21 (20): 1–19. https://doi.org/10.3390/ s21206744.

3. Ragheb, Ayat, Amr Gody, and Tarek Said. 2021. "Comparative Study of Different Types of RNN in Speech Classification. Egyptian J Langu Engineer.2022;8(1):1–16.https://doi.org/10.21608/ejle.2021. 45203.1014.

4.Abdelhamid A A, El Sayed M, El-KenawyBandar A, Ghada M, Amer M Y, Abdelkader A Iim, and Marwa M E. "Robust Speech Emotion Recognition Using CNN+LSTM Based on Stochastic Fractal Search Optimization Algorithm." *IEEE Access.* 2022; 10:49265–84. https://doi.org/ 10.1109/ACCESS.2022.3172954.

5. Vekkot S. Deepa Gupta, Mohammed Zakariah, and Yousef Ajami Alotaibi. 2020. Emotional Voice Conversion Using a Hybrid Framework with Speaker-Adaptive DNN and Particle-Swarm-Optimized Neural Network." *IEEE Access* 2020;8:74627–47. https://doi.org/10.1109 /ACCESS. 2020.2988781.

6. Dey, Arijit, Soham Chattopadhyay, Pawan Kumar Singh, Ali Ahmadian, Massimiliano Ferrara, and Ram Sarkar. A Hybrid Meta-Heuristic Feature Selection Method Using Golden Ratio and Equilibrium Optimization Algorithms for Speech Emotion Recognition. IEEE Access. 2020; 8:200953–70. https://doi.org/10.1109/ACCESS.2020.3035531.

7. Alsayadi H A, Abdelaziz A, Abdelhamid I H and Zaki T F. "Arabic Speech Recognition Using End-to-end Deep Learning." *IET Signal Processing,* 2021;15 (8):521–34. https://doi.org /10. 1049/sil2.12057.

8.Alsobhani A, Hanaa M A, Alabboodi, and Mahdi H. "Speech Recognition Using Convolution Deep Neural Networks. J Physics. 2021. Conference Series 1973 (1). https://doi.org/10.1088/ 1742-6596 /1973 /1/ 012166.

9. Dua S, Sethuraman S K, Albagory Yasser, Rajakumar R, Ankur D, Rajesh S, Mamoon R, Anita G, Sultan S. Alshamrani, and Alghamdi A S 2022. Developing a Speech Recognition System for Recognizing Tonal Speech Signals Using a Convolutional Neural

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a1

Network." *Applied Sciences (Switzerland)* . Alghamdi.;12 (12). https://doi.org/10.3390/app12126223.
10.Yuan J, Chenxiao Li, Qiao W, Ying H, Jialinqing W, Zhima Z, Jianping H, Jilin F, Xuhui S, and Yali W. Lightning Whistler Wave Speech Recognition Based on Grey Wolf Optimization Algorithm. Atmosphere. 2022; 13 (11). https://doi.org/10.3390/atmos13111828.
11. Alsayadi HA, Al-Hagree S, Alqasemi F A and Abdelhamid A A. "Dialectal Arabic Speech Recognition Using CNN-LSTM Based on End-to-End Deep Learning." *2022 2nd International Conference on Emerging Smart Technologies and Applications, ESmarTA 2022*, no. March 2023. https://doi.org/10.1109/ eSmarTA56775 .2022.9935427.
12. Van, Loan Trinh, Thuy Thi Le Dao, Thanh Le Xuan, and Eric Castelli. 2022. "Emotional Speech Recognition Using Deep Neural Networks." *Sensors* 22 (4). https://doi.org/10.3390/s22041414.
13.Rudregowda, Shashidhar, Sudarshan Patilkulkarni, Vinayakumar Ravi, Gururaj H.L., and Moez Krichen. 2023. "Audiovisual Speech Recognition Based on a Deep Convolutional Neural Network." *Data Science and Management* 7 (1): 25–34. https://doi.org/10.1016/j. dsm. 2023.10.002.
10.1016/j.aej.2020.11.004.
14. Thomas, Samuel, Hong Kwang J. Kuo, George Saon, Zoltán Tüske, Brian Kingsbury, Gakuto Kurata, Zvi Kons, and Ron Hoory. 2021. "RNN Transducer Models for Spoken Language Understanding." *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2021-June:7493–97. https://doi.org/ 10.1109/ . 2021.9414029.
15. Abdelfatah H, Ismail S, and Mohamed B A. COVID-19 Detection System Using Recurrent Neural Networks. Proceedings of the 2020 IEEE International Conference on Communications, Computing, Cybersecurity and Informatics, CCCI 2020;1–5.https://doi.org/10.1109/ CCCI49893. 2020.9256562.
16. Zhou Wei, Zuoyun Z, Ralf S, and Hermann N. On Language Model Integration for Rnn Transducer Based Speech Recognition." *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2022-May (Ilm): 8407–11.
Alsayadi H, Abdelhamid A, Hegazy I and Taha. Z. Data Augmentation for Arabic Speech Recognition Based on End-to-End Deep Learning." *I*nternational Journal of Intelligent *Computing and Information Sciences* 2021;21 (2): 50–64. https://doi.org/10.21608/ijicis. 2021.73581.1086.
17. Oruh J, Serestina V, and Adekanmi A. Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition." *IEEE Access* 2022;10:30069–79. https://doi.org/10.1109/ ACCESS. 2022 .3159339.
18. Daneshfar F and Seyed J K. Speech Emotion Recognition Using Discriminative Dimension Reduction by Employing a Modified Quantum-Behaved Particle Swarm Optimization Algorithm." *Multimedia Tools and Applications*. 2019;79 (1–2):1261–89. https:// doi.org/10.1007/s11042-019-08222-8.
19. Rajasekhar B, Kamaraju M, and Sumalatha V. A Novel Speech Emotion RecognitionModel Using Mean Update of Particle Swarm and Whale Optimization-Based Deep Belief Network." *Data Technologies and Applications* . 2020;54 (3): 297–322. https://doi.org/10.1108/DTA-07-2019-0120.
20. Gomathy M. 2020. Optimal Feature Selection for Speech Emotion Recognition Using Enhanced Cat Swarm Optimization Algorithm." *I*nternational Journal of Speech Technology 24 (1): 155–63. https://doi.org/10.1007/s10772-020-09776-x.
21. Singh A. Speech Emotion Recognition Using Enhanced Cat Swarm Optimization Algorithm.International Journal of Information Technology. 2020; 6 (5): 2023–34.
22. Rajasekhar, B., M. Kamaraju, and V. Sumalatha. 2019. "Glowworm Swarm Based Fuzzy Classifier with Dual Features for Speech Emotion Recognition." *Evolutionary Intelligence* 15 (2): 939–53. https://doi.org /10.1007/s12065-019-00262-1.
23. Liangpan, Ye, and He Tao. 2021. HMM Speech Recognition Study of an Improved Particle Swarm Optimization Based on Self-Adaptive Escape (AEPSO)." *IOP Conference Series: Earth and Environmental Science* 634 (1). https://doi.org/10.1088/1755-1315/634/1/012074.
24. Zhang Zicheng. 2020. "Speech Feature Selection and Emotion Recognition Based on Weighted Binary Cuckoo Search." *Alexandria Engineering Journal* 60 (1): 1499–1507. https://doi.org/
25. Rajeshwar J. Hybrid Particle Swarm Optimization-Gravitational Search Algorithm Based Deep Belief Network: Speech Emotion Recognition.Journal of Computational Mechanics, Power System and Control 4 2020; (3): 23–31. https://doi.org/10.46253/jcmps.v4i3.a4.