https://doi.org/10.31436/iiumej.v26i1.3386

GLOBAL-LOCAL SELF-ATTENTION BASED LONG SHORT-TERM MEMORY WITH OPTIMIZATION ALGORITHM FOR SPEAKER IDENTIFICATION

PRAVIN MAROTRAO GHATE[1*], BHAGVAT D. JADHAV[1], SHRIRAM SADASHIV KULKARNI[2], PRAVIN BALASO CHOPADE[3], PRABHAKAR N. KOTA[3]

[1]*Department of Electronics & Telecommunication, JSPM's Rajarshi Shahu College of Engineering, Tathawade, Pune, India*

[2]*Department of Information Technology, Sinhgad Academy of Engineering, Pune, India*

[3]*Department of Electronics & Telecommunication, M.E.S College of Engineering, Pune, India*

**Corresponding author: pmghate@gmail.com*

(Received: 30 July 2024; Accepted: 8 October 2024; Published online: 10 January 2025)

ABSTRACT: Speaker identification (SI) involves recognizing a speaker from a group of unknown speakers, while speaker verification (SV) determines if a given voice sample belongs to a particular person. The main drawbacks of SI are session variability, noise in the background, and insufficient information. To mitigate the limitations mentioned above, this research proposes Global Local Self-Attention (GLSA) based Long Short-Term Memory (LSTM) with Exponential Neighborhood – Grey Wolf Optimization (EN-GWO) method for effective speaker identification using TIMIT and VoxCeleb 1 datasets. The GLSA is incorporated in LSTM, which focuses on the required data, and the hyperparameters are tuned using the EN-GWO, which enhances speaker identification performance. The GLSA-LSTM with EN-GWO method acquires an accuracy of 99.36% on the TIMIT dataset, and an accuracy of 93.45% on the VoxCeleb 1 datasets, while compared to SincNet and Generative Adversarial Network (SincGAN) and Hybrid Neural Network – Support Vector Machine (NN-SVM).

ABSTRAK: Pengenalpastian pembicara (Speaker Identification, SI) melibatkan pengenalan pembicara daripada kumpulan pembicara yang tidak dikenali, manakala pengesahan pembicara (Speaker Verification, SV) menentukan sama ada sampel suara tertentu milik seseorang individu. Kekurangan utama dalam SI ialah variasi sesi, bunyi latar belakang, dan maklumat yang tidak mencukupi. Untuk mengatasi kekangan tersebut, kajian ini mencadangkan kaedah Global Local Self-Attention (GLSA) berasaskan Long Short-Term Memory (LSTM) dengan Pengoptimuman Grey Wolf Jiranan Eksponen (EN-GWO) bagi pengenalpastian pembicara yang berkesan menggunakan set data TIMIT dan VoxCeleb 1. GLSA digabungkan dalam LSTM yang memberi tumpuan pada data yang diperlukan, manakala parameter hiper ditala menggunakan EN-GWO untuk meningkatkan prestasi pengenalpastian pembicara. Kaedah GLSA-LSTM dengan EN-GWO mencapai ketepatan 99.36% pada dataset TIMIT dan ketepatan 93.45% pada dataset VoxCeleb 1, berbanding dengan SincNet dan Generative Adversarial Network (SincGAN) serta Hybrid Neural Network – Support Vector Machine (NN-SVM).

KEYWORDS: *Exponential Neighborhood – Grey Wolf Optimization, Global-Local Self-Attention, Long Short-Term Memory, Mel-frequency Cepstral Coefficients, Speaker Identification.*

https://doi.org/10.31436/iiumej.v26i1.3386

1. INTRODUCTION

Recently, a wide variety of speech-related applications have been introduced, and many of the applications are highly convenient for communication with human-human and human-machine [1]. Every human has a unique voiceprint that reflects the specific acoustics of its voice-producing organs and their patterns of speaking [2-4]. The speech recognition technique is classified into two types: speaker identification (SI), which identifies the speaker from a set of known speakers, and speaker verification (SV), which determines whether a given speaker matches a particular identity of a person [5, 6]. Both verification and identification need a fixed phrase or sentence of agnostic to be attained through the speaker [7]. When the speakers and the text are predetermined, it is called a text-dependent system. Conversely, when speakers can speak freely without fixed constraints, it is a text-independent system [8]. Utilizing the voice-recognition technique effectively enhances the effectiveness of the minutes-development task. The ability to automatically identify speakers in transcripts greatly enhance the overall effectiveness of the minute process [9].

The performance degradation of SI in extra speech is caused by variance among emotion classes in enrolling data that represents training and test information for every speaker in the SI method [10, 11]. The SI method enhances its performance by utilizing speech data that belongs to the same emotional phase in both training and testing [12]. However, gathering balanced speech data that covers different emotional states for every speaker is difficult due to huge costs and consumption of time [13-15]. The main drawbacks of SI are session variability, noise in the background, and lack of enough information [16]. Significant differences often occur between training and testing sessions when training is performed in a clean environment, whereas testing is performed in a noisy environment. The noise in the background is another drawback that affects the accuracy of SI. Recently, deep learning methods like spatially dependent CNN have been utilized for audio and image classification fields for their capacity to extract robust features. The main drawbacks of SI are session variability, noise in the background, and insufficient information. The attention mechanism is incorporated in the DL-based algorithm, which focuses much on required data, and the hyperparameters are tuned by using the optimization algorithm that enhances speaker identification performance. The main contributions of the research are given as follows:

- Mel-frequency Cepstral Coefficients (MFCC) and Zero Crossing Rate (ZCR) techniques are utilized for feature extraction to differentiate the features for enhancing the classification.
- The Global Local Attention mechanism is utilized in LSTM to concentrate on essential features for speaker identification that enhance classification performance.
- Then, the hyperparameters of LSTM are tuned using the Exponential Neighborhood—Grey Wolf Optimization (EN-GWO) algorithm, which enhances classification performance.

This research is organized as follows: Section 2 provides a literature review, Section 3 provides details of the proposed methodology, Section 4 provides results and a discussion of the proposed method, and the conclusion is in Section 5.

2. LITERATURE REVIEW

Wei et al. [17] suggested SincNet and Generative Adversarial Network (SincGAN) methods for speaker identification using speech signals. Unlike other techniques, the implemented method utilized feature recognition of standard hand-crafted. The generator in GAN was used to reconstruct input samples to improve the count of training samples, while the discriminator was used for SI classification. The multiple-scale SincNet layer, depending

on 3 bespoke filter banks, was updated to capture low-level speech representation. The implemented method recognized the raw waveform of the input and allowed speaker identification by a smaller number of training utterances. However, the implemented method obtained speech data from users to extract sufficient data represented by speaker features, which was difficult because of the constraints of the practical environment. However, the implemented method was more time-consuming.

Karthikeyan et al. [18] introduced a Hybrid Neural Network–Support Vector Machine (NN-SVM) method for speaker identification using speech signals. Integrating the NNSVM structure involved passing the output of the Artificial Neural Network (ANN) to an SVM classifier to obtain the results. The introduced method integrated NN-SVM to classify multiple class imbalanced speech information, which obtained the respective advantages of machine learning and SVM structure with a reduced loss function. However, the introduced method had additional costs for the server.

Shahamiri [19] presented an Enhanced Multi-Active Learner (EMAL) method for speaker identification using speech signals. The presented method distributed the difficulty of the learning task between a learner's array with an effective SI technique in speakers, which were found to depend on one sound segment of voice biometrics. The datasets used for evaluation were VoxCeleb 1 and TIMIT. However, the presented method was too limited for lower frequency parts, unable to capture resonant peaks.

Barhoush et al. [20] suggested two parts of techniques, Fully Connected – Deep Neural Network (FC-DNN) and MFCC, depending on the features of Shuffled MFCC (SHMFCC), along with their Difference Shuffled MFCC (DSHMFCC) methods for speaker identification using speech signals. The FC-DNN method effectively localized and identified active speakers separately with huge accuracy. However, the suggested method was vulnerable to security issues.

Gaurav et al. [21] implemented Mask Region-based CNN (Mask R-CNN) technique parameter optimization using the Hosted Cuckoo Optimization (HCO) method for speaker identification using speech signals. From the input speech signal, four kinds of features were extracted: MF Differential Power (MFDPCC), Gamma tone Frequency (GFCC), Power Normalized (PNCC), and Spectral entropy to enhance signal strength. Next, the speaker's ID was classified through Mask R-CNN, with its parameters undergoing optimization through the HCO algorithm.

Al Dulaimi et al. [22] introduced a two-dimensional discrete Multi-Wavelet Transform (2D-DMWT) with deep learning neural network methods for speaker identification using speech signals. The introduced method was dependent on essential sampling pre-processing, which utilized a filter through Geronimo, Hardian, and Massopust, known as GHM. The 2D-DMWT was assigned to acquire discriminant features from the speech signal with minimized dimensions of the speech signal during the feature selection phase. Finally, CNN was utilized to classify SI.

3. PROPOSED METHODOLOGY

This research proposes a Global-Local Self-Attention (GLSA) based Long Short-Term Memory (LSTM) with EN-GWO for speaker identification using speech signals. The datasets utilized for research are TIMIT and VoxCeleb 1, alongside the pre-processing techniques, namely, median filter, pre-emphasis, and de-emphasis. Feature extraction is performed by MFCC, ZCR, and pitch extraction, while LSGA-based LSTM identifies the speaker with EGWO. Figure 1. illustrates the overall process of the proposed algorithm.

https://doi.org/10.31436/iiumej.v26i1.3386

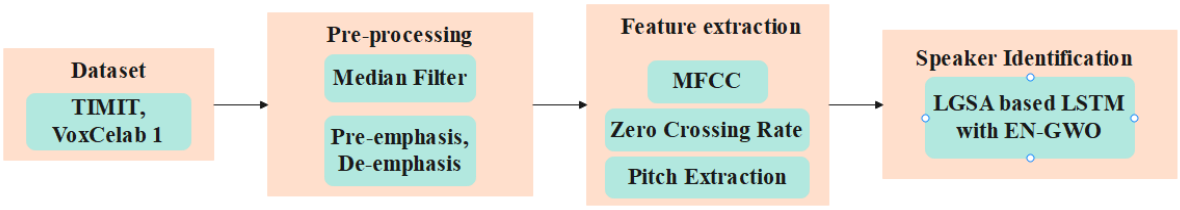

Figure 1. Overall Process of Proposed Algorithm

3.1. Dataset

The datasets used for research are TIMIT [23] and VoxCeleb 1 [24], which are audio signal datasets. These audio signal datasets are used for speaker identification. The brief details of these datasets are described below.

3.1.1. TIMIT dataset

The TIMIT is an Acoustic-Phonetic Continuous Speech Corpus that has 6300 speech data of quality, clean speech samples acquired from 630 speakers, and 10 speeches from each speaker are generated. Of 630 speakers, 192 are females, and 438 are males from 8 main areas of the USA.

3.1.2. VoxCeleb 1 dataset

The VoxCeleb 1 dataset has 153K gender-balanced utterances from 1251 speakers taken from videos on YouTube. This dataset includes 352 hours of speech and 55% of male samples of speech gathered from speakers of a varied range of ethnicities, professions, and ages. There is an average value of 116 speeches for every speaker, and the mean length of every sample is 8.2 seconds. These two datasets are used in the pre-processing techniques to reduce noise and boost the frequency phase of signals.

3.2. Pre-processing

Pre-processing is a vital phase in speaker identification. The methods utilized are median filter, pre-emphasis, and de-emphasis. The median filter reduces noise in the dataset, with pre-emphasis boosting the high-frequency phase while de-emphasis cutting the signal's low-frequency phase.

3.2.1. Median Filter

The median filter is a non-linear digital filtering method utilized for removing noise from dataset signals. Reducing noise in the pre-processing phase enhances the signals for feature extraction. The median filter eliminates the impact of input noise values with excessively high magnitudes. The median filter maintains the edge integrity, which is mainly used in speaker identification, where the boundary data are essential. The mathematical formula for reducing noise using a median filter is given in Eq. (1),

$$y[i] = \frac{1}{M}\sum_{j=0}^{M-1} x[i+j] \tag{1}$$

where, $x[]$ represents input signal, $y[]$ represents output signal, and M represents the number of signals used in the median. The median filter eliminates the effect of input noise with large magnitudes.

https://doi.org/10.31436/iiumej.v26i1.3386

3.2.2. Pre-emphasis and De-emphasis

The pre-emphasis is processed by boosting the signal's high-frequency phase, compensating for a high-frequency loss in the cable. The de-emphasis is processed by cutting the less-frequency phase in a signal coupled with maximized transmit voltage. Pre-emphasis is used in speech processing to increase high-frequency elements of speech signals and minimize fewer frequency elements. De-emphasis is utilized in speech processing to minimize the magnitude of specific (high) frequencies concerning other (less) frequencies and enhance the signal-to-noise ratio. The Pre-emphasis and De-emphasis balance the signal's frequency spectrum, enhance Signal Noise Ratio (SNR), and make it essential for high-frequency features. The pre-processed signal is given as input to the feature extraction phase to extract the features required for speaker identification.

3.2. Feature Extraction

Feature extraction is a process of identifying a compact and meaningful set of features from raw data that improves the classifier's efficiency. In this research, feature extraction is utilized to extract features from pre-processed signals for effective speaker identification. MFCC, ZCR, and pitch extraction techniques are utilized for feature extraction in pre-processed signals.

3.3.1. Mel-frequency Cepstral Coefficients

The MFCC gives spectral speech data and characterizes a perception of human hearing. MFCC has characteristics of huge accuracy and stable recognition. At a similar time, the benefits of MFCC included their ability to provide high-fidelity representations, making them a potential candidate in the speech recognition field. The MFCC gives a meaningful representation of the audio signal and captures significant data about the timbre and speech sounds. The cepstral coefficients extracted by MFCC are efficient in recognizing the speaker, captured content, and individual characteristics of the speaker's voice. Figure 2. provides the flow of MFCC.

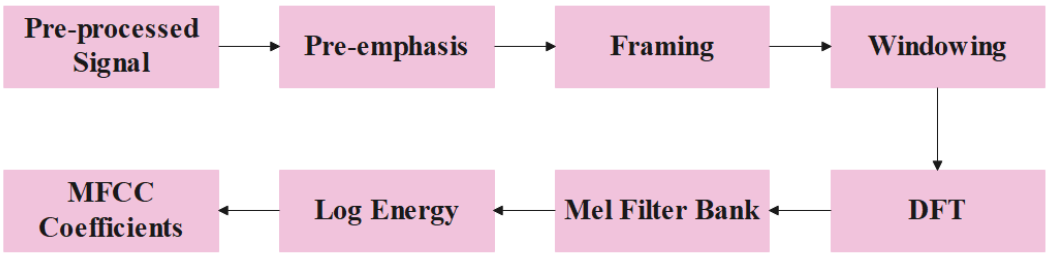

Figure 2. Provides the flow of MFCC

In the MFCC extraction procedure, pre-emphasis is utilized on normalized speech signals to reduce noise in the raw speech signals. The filtered signal is separated into 40ms frames with a 50% of frameshift. For 4s speech signals, the whole 199 frames are produced as a 40ms frame width. The hamming window with $\alpha = 0.46$ and N represents the count of samples for each frame length of 30ms collected nearest to the frequency elements together, and the mathematical formula is given as Eq. (2),

$$H(n) = (1 - \alpha) - \alpha \times cos\left(\frac{2\pi n}{(N-1)}\right),\ 0 \leq n \leq N - 1 \quad (2)$$

https://doi.org/10.31436/iiumej.v26i1.3386

Next, the Discrete Fourier Transform (DFT) is utilized for converting speech signal time-domain to frequency domain ($X(k)$), and the mathematical formula is given as Eq. (3). Eq. (4) represents the power spectrum of DFT that represents characteristics of the vocal tract. Next, the signal is processed by triangular filter banks of Mel-frequency ($\nabla_m(k)$) the mathematical formula of Eq. (5) generates the conversion from linear to mel frequency to give speech-hearing perceptual data.

$$X(k) = \sum_{n=0}^{N-1} x(n) \times H(n) \times e^{-j2\pi nk/N}, 0 \leq n, k \leq N-1 \quad (3)$$

$$X_k = \frac{1}{N}|X(k)|^2 \quad (4)$$

$$ET_m = \sum_{k=0}^{k=1} \nabla_m(k) \times X_k; \quad m = 1,2,\dots,M \quad (5)$$

Next, the Discrete Cosine Transform (DCT) of the log-filter bank energy signal gives L count of cepstral coefficients with its numerical formula is given as Eq. (6)

$$MFCC_i = \sum_{m=1}^{M} log_{10}(ET_m) \times cosj\left((m+0.5)\frac{\pi}{m}\right), j = 1,2,\dots,L \quad (6)$$

The extracted features significantly characterize the transition in speaker identification. The MFCC provides energy, MFCC coefficients, and significant derivative features for characterizing speech signals.

3.3.2. Zero Crossing Rate

The ZCR generates a transition signal over a zero line, representing the speech signal's noise metrics. The ZCR is an effective feature extraction technique that supports the differentiation between various types of sounds or energy variations in audio. The sign function gives a value of 1 for positive and 0 for negative sample amplitudes across the time frame (t). The mathematical Eq. (7) generates the execution of ZCR in the time domain.

$$ZCR_t = \frac{1}{2}(\sum_{n=1}^{N} sign(x[n]) - sign(x[n-1])) \quad (7)$$

where, $sign(x[n])$ represents the sign function and t represents the time frame. The ZCR counts how many times the speech signal crosses the zero axis, indicating modifications or occurrences in some sounds. The ZCR method produces less noise during the switching process.

3.3.3. Pitch Extraction by Synthesis based method

In synthesis-based pitch extraction, the speech signal is considered a result of an all-pole filter $H(z)$ that is excited through the period impulse train or random noise. This method captures the essential signal frequency essential to identify the intonation patterns. Therefore, $S(f)$ power spectrum of the input speech segment is acquired through DFT, and the mathematical formula for decomposed is given in Eq. (8),

$$S(f) = H(f)E(f) \quad (8)$$

where, $E(f)$ represents the power spectrum in the excited signal. To synthesize the power spectrum, the input speech segment is considered as voice, and their frequency is hypothesized $\check{F}_o$. Transfer function H(z) of every-pole filter is evaluated from the input speech segment through linear prediction analysis. The numerical formula for the synthesized power spectrum is given in Eqs. (9) and (10).

$$S(f, \bar{F}_o) = H(f)E(f, \bar{F}_o) \quad (9)$$

https://doi.org/10.31436/iiumej.v26i1.3386

where,

$$\bar{E}(f,\bar{F}_o) = \begin{cases} 1, & for\ f = \bar{F}_o, 2\bar{F}_o, 3\bar{F}_o \\ 0, & otherwise \end{cases} \quad (10)$$

The amount of power spectra is synthesized for various scores of $\bar{F}_o$ with hypothesized frequency compared to S(f). The dissimilarity among Average Magnitude (AM) is compared to measure the variance between the input and synthesized power spectrum. The mathematical expression is in Eq. (11).

To compare the dissimilarity between the two spectra, the Average Magnitude (AM) variance between the input power spectrum and the synthesized power spectrum is calculated. This helps quantify differences in spectral characteristics.

$$D(\bar{F}_o) = \frac{1}{F_f}\int_0^{F_f} |d(f,\bar{F}_o)|\, df \quad (11)$$

where, F_f represents the folding frequency and $d(f,\bar{F}_o)$ represents variance among two spectra at f frequency and its mathematical formula is given as Eq. (12),

$$d(f,\bar{F}_o) = logS(f) - log\,\bar{S}(f,\bar{F}_o) \quad (12)$$

Rather than using power spectra $S(f)$ and $\bar{S}(f,\bar{F}_o)$, log power is utilized in the above equation for defining variance $d(f,\bar{F}_o)$. The value of fundamental frequency, which provides less AM variance among the synthesized spectrum and input spectrum, is selected as an evaluated score of significant frequency. This is performed to remove the effect of format architecture on pitch estimation. The extracted features are significant in speaker identification and are given as input to LSTM for identifying speakers. Finally, seven features of mean, median, standard deviation, variance, and so on are extracted. Various features give spectral modifications in speech signals because of modifications in metrics and intonation in speech signals. The extracted features are given as input to the neural network for identifying speakers.

3.3. Speaker Identification using GLSA-LSTM with EN-GWO

The extracted features are given as input to Long-Short-Term Memory (LSTM), which utilizes Global-Local Self-Attention (GLSA) to concentrate on the significant parts of the feature. The weight of LSTM is updated using the proposed Exponential Neighborhood—Grey Wolf Optimization (EN-GWO). Below are detailed explanations of LSTM, GLSA, and the proposed EN-GWO algorithms.

3.3.1. Long Short-Term Memory (LSTM)

The architecture of LSTM is represented in Figure 3, where the controlled state of information is considered at every moment in the whole LSTM using a structure known as the gate for attaining the learning effect. C_t represents state information of the LSTM unit at the time t, f_t represents forget gate at the time t, i_t represents the input gate at the time t, $\tilde{C}_t$ represents present-moment information, o_t represents the output gate at the time t, $tanh$ represents tangent activation function and σ represents the sigmoid activation function.

https://doi.org/10.31436/iiumej.v26i1.3386

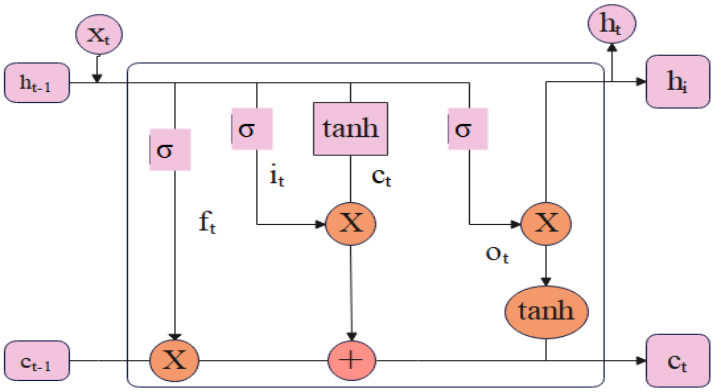

Figure 3. Architecture of LSTM

The mathematical formula for every gate is given from Eqs. (13) – (18),

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{13}$$

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{14}$$

$$\tilde{C}_t = tanh\ (W_{\tilde{c}} \cdot [h_{t-1}, x_t] + b_{\tilde{c}} \tag{15}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{16}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{17}$$

$$h_t = o_t \cdot tanh C_t \tag{18}$$

where, $W_f, W_i, W_{\tilde{c}}\ and\ W_0$ represent weight matrices respective to every module, $b_f, b_i, b_{\tilde{c}}\ and\ b_o$ represent bias and mathematical formulas for tanh( ) and σ is given as in Eqs. (19) and (20),

$$tanhx = (e^x - e^{-x})/(e^x + e^{-x}), \tag{19}$$

$$\sigma(x) = 1/(1 + e^{-x}) \tag{20}$$

At last, the output layer is dependent on h_t by a fully connected layer for obtaining the predicted value y_t, given as Eq. (21),

$$y_t = \sigma\left(W_y \cdot h_t + b_y\right) \tag{21}$$

where, W_y represents the weight matrix and b_y represents bias term. The LSTM network provides much more accurate prediction using the features of past time points of time series. The initial position of the past phase is called currently, representing a phase of the familiar signal.

3.3.2. Global-Local Self Attention

The self-attention mechanism has global dependencies but struggles to capture local data for utterances. This research utilizes a global-local self-attention mechanism, which enhances the ability to capture local features while retaining the ability of the method's long-term dependencies. Figure 4 represents the flow of Global-Local Self Attention.

https://doi.org/10.31436/iiumej.v26i1.3386

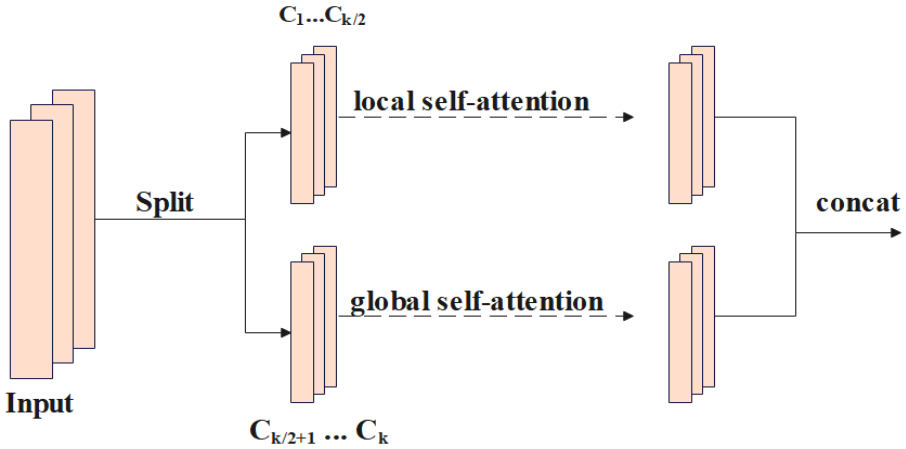

Figure 4. The Flow of Global-Local Self-Attention Mechanism

The SA mechanism was initially introduced with LSA, where every attention head included a sliding window of a specified size. As the next half channel is introduced, GSA is utilized without a sliding window. Similar to the actual SA mechanism, input features $X \in R^{C\times T}$ are linearly converted to K attention head, and next, every attention head processes local and global SA. In GLSA, the non-overlapping sliding window partitions X into $[X_1, \dots, X_N]$ and each segment has the same window size w. It is considered that matrices of query, key, and value of kth attention head contains dimension d_k. The mathematical formula for LSA for kth head are given in Eqs. (22) – (24),

$$X = [X_1, X_2, \dots, X_N],\ N = \frac{T}{w} \tag{22}$$

$$Y_i^k = Attention\left(X_i W_k^Q, X_i W_k^K, X_i W_k^V\right) \tag{23}$$

$$LocalAttention_k(X) = \left[Y_1^k, Y_2^k, \dots, Y_i^k\right] \tag{24}$$

where, $W_k^Q, W_k^K, W_k^V \in R^{C\times d_k}$ represents linear projection parameter matrix of query, key, and value in kth attention head, d_k represents set to C/K. K attention head is divided correspondingly into two distinct sets. The K is generally an even number, with the attention head equally separated into two sets. The initial set of attention heads performs LSA, and the next set of attention heads process global self-attention. The mathematical formula for output kth head is given in Eq. (25),

$$head_k = \begin{cases} LocalAttention_k(X), & k = 1, \dots, \frac{K}{2} \\ GlobalAttention_k(X) & k = \frac{K}{2} + 1, \dots, K \end{cases} \tag{25}$$

At last, the outcomes of two attention mechanisms are combined as input of LSTM and represented as GL – Attention(X), with the mathematical formula given as Eq. (26).

$$GL - Attention(X) = cat(head_1, \dots, head_k)W \tag{26}$$

where, $W \in R^{C\times C}$ represents a projection matrix projecting self-attention output to the target result dimension. The main plan is to separate attention heads into 2 various sets and process GLSA in parallel. That enables local attention within the global attention process, allowing global data to interact with local data in a structured manner.

https://doi.org/10.31436/iiumej.v26i1.3386

3.3.3. Exponential Neighborhood Grey Wolf Optimization

Exponential Neighborhood Grey Wolf Optimization (EN-GWO) emulates grey wolves' group hunting behavior and leadership hierarchy. This mathematical method models the prey's location as the optimal solution in the search space, while the wolf's current position represents a potential solution. In this approach, the positions of the alpha, beta, and delta wolves are denoted as α, β, and γ, respectively. The EN-GWO algorithm involves four significant phases: encircling, hunting, attacking, and seeking prey. During the hunting phase, the wolves encircle their prey, which is mathematically modeled using the equations provided in Eqs. (27)–(30).

$$D = \left|C - X_p(t) - X_i(t)\right| \tag{27}$$

$$X_i(t+1) = X_p(t) - AD \tag{28}$$

$$C = 2a \times rand_1 - a \tag{29}$$

$$A = 2rand_2 \tag{30}$$

where, $X_i(t)$, $X_p(t)$ represents ith grey wolf position vector and prey at t iteration, $A\ and\ C$ represents coefficient vectors, $rand_1$ and $rand_2$ represents random vectors in [0,1], a represents the vector whose values are minimized from iteration 2 – 0. In the hunting process, all wolves are the same with probable prey positions, and from that, three good agents are developed with their mathematical formula given in Eqs. (31) – (37).

$$D_\alpha = |C_1 X_\alpha - X_i| \tag{31}$$

$$D_\beta = \left|C_2 X_\beta - X_i\right| \tag{32}$$

$$D_\gamma = \left|C_3 X_\gamma - X_i\right| \tag{33}$$

$$X_1 = X_\alpha - A_1 . D_\alpha \tag{34}$$

$$X_2 = X_\beta - A_2 . D_\beta \tag{35}$$

$$X_3 = X_\gamma - A_3 . D_\gamma \tag{36}$$

$$X_{iGWO}(t+1) = \frac{X_1 + X_2 + X_3}{3} \tag{37}$$

Linear minimization in adaptive feature a across iterations helps balance the exploitation and exploration of attacking methods. Another adaptive parameter manages exploitation and exploration A that considers random values ranging from $[-2r, -2r]$. The exploitation develops the local optimum, which yields an optimum and lower quality. The default GWO represents half of their cycles for exploitation and the next half for exploitation. The diversity, lack of population, and premature convergence are other difficult factors of traditional GWO. To overcome this limitation, the research proposes an Exponential Neighborhood Grey Wolf Optimization by enhancing two phases of default GWO. The mathematical formula for introducing the exponential decline function for managing adaptive parameter $\vec{a}$ value is given in Eq. (38). Linear minimization in adaptive feature selection across iterations helps balance between exploitation and exploration in the attacking methods.

$$a = 2\left(1 - \frac{t^2}{T^2}\right) \tag{38}$$

where, t represents the present iteration and T represents the whole iteration. The neighborhood-based searching process is introduced by combining an individual wolf-hunting strategy with a global hunting strategy. Multi-neighbors are utilized to expand the search space. The initially updated location of the present wolf $X_{iGWrewrO}(t)$ is measured by Eq. (33). Next,

$R_i(t)$ is executed by Euclidean distance among the present location $X_i(t)$ and $X_{iGWO}(t)$ is given as Eq. (39).

$$R_i(t) = X_i(t) - X_{iGWO}(t) \tag{39}$$

Respective to the radius, the neighbors of the present wolf $X_i(t)$ are measured using below Eq. (40) with wolf $X_j(t)$ distance from the present wolf from the radius value.

$$N_i(t) = \left\{X_j(t) | D_i\left(X_j(t), X_i(t)\right) \leq R_i(t)\right\} \tag{40}$$

The next new position $X_{NL}(t)$ depends on the position of neighbors, as developed by Eq. (41),

$$X_{NL}(t) = X_i(t) + rand \times \left(X_n(t) - X_r(t)\right) \tag{41}$$

where, $X_n(t)$ represents the randomly chosen neighbor N(t) and $X_r(t)$ represents the randomly chosen wolves through the population. At last, good location within X_{iGWO} and X_{NL} with respect to its fitness value for further iteration is given as Eq. (42),

$$X_i(t+1) = \begin{cases} X_{iGWO}, iff(X_{iGWO}) > f(X_{NL}) \\ X_{NL}, \qquad\qquad otherwise \end{cases} \tag{42}$$

While utilizing a global-local self-attention mechanism at various gates, LSTM gives enhanced outcomes. When utilizing the EN-GWO algorithm for LSTM, every position vector of the wolf represents the weight of LSTM. After completing a set of iterations, the position vector of a wolf with the least cost in the EN-GWO algorithm is selected as the final weight for the trained network. At last, the robustness and accuracy are enhanced with the speaker's identity.

4. EXPERIMENTAL RESULTS

The GLSA-LSTM with EN-GWO algorithm is simulated on a Python environment with system configurations of an i5 processor and 16GB RAM. The performance metrics used for evaluating the GLSA-LSTM with the EN-GWO method are accuracy, precision, recall, and f1-score. The numerical expression for performance metrics is given in Eqs. (43) – (46).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{43}$$

$$Precision = \frac{TP}{TP+FP} \tag{44}$$

$$Recall = \frac{TP}{TP+FN} \tag{45}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{46}$$

where, the TP – True Positive, FP – False Positive, TN – True Negative and FN – False Negative.

4.1. Quantitative and Qualitative Analysis

The performance of the GLSA-LSTM with the EN-GWO technique is analyzed using performance measures of accuracy, precision, recall, and f1-score. Different tables and graphs show the proposed algorithm's performance.

https://doi.org/10.31436/iiumej.v26i1.3386

Table 1. Performance of proposed algorithm with 2 datasets

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
TIMIT	99.36	99.12	97.62	98.47
VoxCeleb 1	93.45	92.78	92.04	91.58
Average	96.40	95.95	94.83	95.02

Table 1 represents the performance of the proposed method with two datasets. The developed method attains an accuracy of 99.36%, precision of 99.12%, recall of 97.62%, and f1-score of 98.47% on the TIMIT dataset. The GLSA attains an accuracy of 93.45%, precision of 92.78%, recall of 92.04, and f1-score of 91.58%, with the mean of the two datasets represented.

Table 2. Performance of Global-Local Self-Attention Mechanism with 2 datasets

Attention Mechanism	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SE	90.40	90.77	89.83	86.18
CAM	91.39	91.02	90.17	87.27
SAM	92.24	91.82	90.94	88.72
SA	93.58	92.36	91.72	89.37
GLSA	95.67	93.29	92.15	90.48

Table 2 describes the performance of the GLSA mechanism with two datasets. The existing attention mechanisms considered for evaluation are Squeeze and Excitation (SE), Channel Attention Module (CAM), Spatial Attention Module (SAM), and Self Attention (SA). Compared to existing mechanisms, the utilized attention mechanism attains an accuracy of 95.67%, precision of 93.29%, recall of 92.15%, and F1-score of 90.48%.

Table 3. Performance of Proposed EN-GWO algorithm with 2 datasets

Optimization algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CSO	92.25	92.48	90.88	89.57
GOA	93.17	93.01	91.03	90.63
WOA	93.99	93.82	92.37	91.03
GWO	95.29	94.61	93.72	92.25
EN-GWO	96.72	95.34	94.28	93.66

Table 3 describes the performance of EN-GWO with 2 datasets. The existing optimization algorithms utilized for evaluation are Cat Swarm Optimization (CSO), Grasshopper Optimization Algorithm (GOA), Whale Optimization Algorithm (WOA), and Grey Wolf Optimization (GWO). The utilized optimization algorithm attains an accuracy of 96.72%, precision of 95.34%, recall of 94.28%, and F1-score of 93.66%, which is better than the existing algorithms.

Table 4. Performance of LSTM neural network with 2 datasets

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
MLP	93.06	92.73	91.63	90.52
ANN	94.16	93.66	92.84	91.03
CNN	95.37	94.25	93.74	92.39
RNN	96.29	95.48	94.21	93.47
LSTM	97.48	96.26	95.37	94.28

https://doi.org/10.31436/iiumej.v26i1.3386

Table 4 describes the performance of LSTM with 2 datasets. The existing Neural networks utilized for evaluation are Multi-Layer Perceptron (MLP), Artificial Neural Network (ANN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN). The utilized neural network attains an accuracy of 97.48%, precision of 96.26%, recall of 95.37%, and F1-score of 94.28%, which is better than existing classifiers.

Table 5. Performance of Proposed GLSA-based LSTM with EN-GWO with 2 datasets

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
GLSA based MLP	91.38	90.37	89.73	90.63
GLSA based ANN	92.17	91.05	90.04	91.85
GLSA based CNN	93.49	92.71	91.74	92.94
GLSA based RNN	94.63	93.84	92.28	93.74
GLSA-based LSTM with EN-GWO	96.40	95.95	94.83	95.02

Table 5 describes the performance of the proposed GLSA-based LSTM with EN-GWO algorithm on two datasets. The existing attention mechanisms utilized for evaluation are GLSA-based MLP, GLSA-based ANN, GLSA-based CNN, and GLSA-based RNN. The utilized optimization algorithm attains an accuracy of 96.72%, precision of 95.34%, recall of 94.28%, and f1-score of 93.66%, outperforming the existing methods.

4.2. Comparative Analysis

The GLSA-LSTM with EN-GWO method is compared to the existing methods: SincGAN [17], Hybrid NN-SVM [18] and EMAL [19]. The GLSA-LSTM with EN-GWO method attains an accuracy of 99.36%, precision of 99.12%, recall of 97.62%, and f1-score of 98.47% on the TIMIT dataset. The attained accuracy of 93.45%, the precision of 92.78%, the recall of 92.04, and the f1-score of 91.58% are all more effective than the existing methods. Combining GLSA, LSTM, and EN-GWO introduces the computational overhead because of the quadratic complexity in GLSA and the sequence nature of LSTM in terms of memory usage and processing power. Table 6 describes the comparative analysis of GLSA-LSTM with the EN-GWO method.

Table 6. Comparative Analysis of GLSA-LSTM with EN-GWO method

Method	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SincGAN [17]	TIMIT	98.98	NA	NA	NA
	LIBRISPEECH	99.27	NA	NA	NA
Hybrid NN-SVM [18]	TIMIT	96.72	99.05	79.75	88.35
	ELSDSR	96.91	98.38	64.88	78.19
EMAL [19]	TIMIT	98.61	NA	NA	NA
	VoxCeleb 1	82.93	NA	NA	NA
Proposed GLSA-based LSTM with EN-GWO	TIMIT	99.36	99.12	97.62	98.47
	VoxCeleb 1	93.45	92.78	92.04	91.58

5. CONCLUSION

The main drawbacks of SI are session variability, noise in the background, and lack of enough information. To mitigate the limitations, this research proposes a GLSA-based LSTM with EN-GWO algorithm for effective speaker identification using TIMIT and VoxCeleb 1 dataset. The Median filter and pre-emphasis and de-emphasis techniques are employed for pre-processing. At the same time, the MFCC and ZCR methods are deployed for feature extraction

to extract the significant features. Next, the speaker identification is performed using GLSA-based LSTM with the EN-GWO algorithm. The proposed method acquires an accuracy of 99.36% on the TIMIT dataset and an accuracy of 93.45% on the VoxCeleb 1 datasets, which is superior to existing methods, SincGAN and NN-SVM. Combining GLSA, LSTM, and EN-GWO introduces the computational overhead because of the quadratic complexity in GLSA and the sequence nature of LSTM in terms of memory usage and processing power. In future work, we will apply the model to real-time applications and work on minimizing the computational overhead.

REFERENCES

[1] Shah SM, Moinuddin M, Khan RA. (2022) A Robust Approach for Speaker Identification Using Dialect Information. Appl. Comput. Intell. Soft Comput., 2022(1):4980920. https://doi.org/10.1155/2022/4980920

[2] Nassif AB, Alnazzawi N, Shahin I, Salloum SA, Hindawi N, Lataifeh M, Elnagar A. (2022) A Novel RBFNN-CNN Model for Speaker Identification in Stressful Talking Environments. Applied Sciences,12(10):4841. https://doi.org/10.3390/app12104841

[3] Dua S, Kumar SS, Albagory Y, Ramalingam R, Dumka A, Singh R, Rashid M, Gehlot A, Alshamrani SS, AlGhamdi AS. (2022) Developing a Speech Recognition System for Recognizing Tonal Speech Signals Using a Convolutional Neural Network. Applied Sciences, 12(12):6223. https://doi.org/10.3390/app12126223

[4] Monir M, Kareem M, El-Dolil SM, Saleeb A, El-Fishawy AS, Nassar MAE, Eldin MZA, Abd El-Samie FE. (2022) Cancelable speaker identification based on cepstral coefficients and comb filters. Int. J. Speech Technol., 25(2):471-492. https://doi.org/10.1007/s10772-021-09804-4

[5] Chen YW, Hung KH, Li YJ, Kang ACF, Lai YH, Liu KC, Fu SW, Wang SS, Tsao Y. (2022) CITISEN: A Deep Learning-Based Speech Signal-Processing Mobile Application. IEEE Access, 10:46082-46099. doi: 10.1109/ACCESS.2022.3153469

[6] Noh K, Jeong H. (2023) Emotion-Aware Speaker Identification with Transfer Learning. IEEE Access, 11:77292-77306. doi: 10.1109/ACCESS.2023.3297715.

[7] Nakamura E, Kageyama Y, Hirose S. (2022) LSTM-based japanese speaker identification using an omnidirectional camera and voice information. IEEJ Trans. Electr. Electron. Eng., 17(5):674-684. https://doi.org/10.1002/tee.23555

[8] De Lima TA, Abreu MCD. (2022) Phoneme analysis for multiple languages with fuzzy-based speaker identification. IET Biom., 11(6):614-624. https://doi.org/10.1049/bme2.12078

[9] Saritha B, Laskar MA, Kirupakaran AM, Laskar RH, Choudhury M, Shome N. (2024) Deep Learning-Based End-to-End Speaker Identification Using Time–Frequency Representation of Speech Signal. Circuits Syst. Signal Process., 43(3):1839-1861. https://doi.org/10.1007/s00034-023-02542-9

[10] Shome N, Saritha B, Kashyap R, Laskar RH. (2023) A robust DNN model for text-independent speaker identification using non-speaker embeddings in diverse data conditions. Neural Comput. Appl., 35(26):18933-18947. https://doi.org/10.1007/s00521-023-08736-1

[11] Al-Karawi KA, Mohammed DY. (2023) Using combined features to improve speaker verification in the face of limited reverberant data. Int. J. Speech Technol., 26:789-799. https://doi.org/10.1007/s10772-023-10048-7

[12] Kuppusamy K, Eswaran C. (2022) Convolutional and Deep Neural Networks based techniques for extracting the age-relevant features of the speaker. J. Ambient Intell. Hum. Comput., 13(12):5655-5667. https://doi.org/10.1007/s12652-021-03238-1

[13] Radha K, Bansal M. (2023) Closed-set automatic speaker identification using multi-scale recurrent networks in non-native children. Int. J. Inf. Technol., 15(3):1375-1385. https://doi.org/10.1007/s41870-023-01224-8

[14] Malek J, Jansky J, Koldovsky Z, Kounovsky T, Cmejla J, Zdansky J. (2022) Target speech extraction: Independent vector extraction guided by supervised speaker identification. IEEE/ACM Trans. Audio Speech Lang. Process., 30:2295-2309.

[15] El Shafai W, Elsayed MA, Rashwan MA, Dessouky MI, El-Fishawy AS, Soliman NF, Alhussan

AA, Abd El-Samie FE. (2023) Optical Ciphering Scheme for Cancellable Speaker Identification System. Computer Systems Science and Engineering, 45(1):563-578. https://doi.org/10.32604/csse.2023.024375

[16] El-Gazar S, El Shafai W, El Banby G, Hamed HFA, Salama GM, Abd-Elnaby M, Abd El-Samie FE. (2022) Cancelable Speaker Identification System Based on Optical-Like Encryption Algorithms. Computer Systems Science and Engineering, 43(1):87-102. DOI:10.32604/csse.2022.022722

[17] Wei G, Zhang Y, Min H, Xu Y. (2023) End-to-end speaker identification research based on multi-scale SincNet and CGAN. Neural Comput. Appl., 35(30):22209-22222. https://doi.org/10.1007/s00521-023-08906-1

[18] Karthikeyan V, Priyadharsini SS, Balamurugan K, Ramasamy M. (2022) Speaker identification using hybrid neural network support vector machine classifier. Int. J. Speech Technol., 25(4):1041-1053. https://doi.org/10.1007/s10772-021-09902-3

[19] Shahamiri SR. (2023) An optimized enhanced-multi learner approach towards speaker identification based on single-sound segments. Multimedia Tools Appl,, 83:24541-24562. https://doi.org/10.1007/s11042-023-16507-2

[20] Barhoush M, Hallawa A, Schmeink A. (2023) Speaker identification and localization using shuffled MFCC features and deep learning. Int. J. Speech Technol., 26(1):185-196. https://doi.org/10.1007/s10772-023-10023-2

[21] Gaurav, Bhardwaj S, Agarwal R. (2023) An efficient speaker identification framework based on Mask R-CNN classifier parameter optimized using hosted cuckoo optimization (HCO). J. Ambient Intell. Hum. Comput., 14(10):13613-13625. https://doi.org/10.1007/s12652-022-03828-7

[22] Al-Dulaimi HW, Aldhahab A, Al Abboodi HM. (2023) Speaker Identification System Employing Multi-resolution Analysis in Conjunction with CNN. International Journal of Intelligent Engineering & Systems, 16(5):350-363. DOI: 10.22266/ijies2023.1031.30

[23] Dataset TIMIT: https://www.kaggle.com/datasets/nltkdata/timitcorpus.

[24] Dataset VoxCeleb 1: https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox1.html.