EASR

Engineering and Applied Science Research

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

Optimized intelligent speech signal verification system for identifying authorized users

Pravin Marotrao Ghate*[1)], Bhagvat D Jadhav[1)], Prabhakar N Kota[2)], Shankar Dattatray Chavan[3)] and Pravin Balaso Chopade[2)]

[1)]Department of Electronics and Telecommunication, JSPM's Rajarshi Shahu College of Engineering, Tathawade, Pune, Maharashtra – 411033, India
[2)]Department of Electronics and Telecommunication, M. E. S. College of Engineering, Pune, Maharashtra – 411001, India
[3)]Department of Electronics and Telecommunication Engineering, Dr. D. Y. Patil Institute of Technology, Pune, Maharashtra – 411018, India

Received 21 April 2023
Revised 29 June 2023
Accepted 26 September 2023

Abstract

Speech processing is today's trending topic in the digital industry for making authentication to keep aware of unauthorized ones. However, analyzing the signal feature using conventional filtering or the neural models is insufficient due to the present signal's several noisy features. Hence, incorporating the different noise elimination filters has maximized the algorithm complexity in verifying the user speech signal. So, the present study built a novel Chimp-based recursive Speech Identification (CbRSI) system for the speech processing domain to verify the authenticated users through the speech signal data. To make signal processing the most straightforward task was activated the filtering function to recognize and neglect the noisy features. Consequently, the filtered audio data is imported as the classification phase input then the features-selecting process is performed. Finally, authenticated users were found and validated the performance by matching the analyzed signal features with the saved audio features. Hence, a novel CbRSI earned the finest user verification exactness score of 98.2%, which is the most satisfactory outcome compared to past studies. Therefore, the implemented solution is the most required framework for verifying authenticated users.

Keywords: Speech processing, Authorized users, Speech verification, Audio signal, Noise filtering

1. Introduction

The speech analysis system is applicable in many authenticated digital sectors for finding authorized and unauthorized users [1]. The objective of a procedure would be to acquire property expertise on one activity and transmit it onto an acceleration program [2]. Noise-resistant voice processing systems are also agnostic of speakers and languages [3]. Specifically, learning unattended audio abstractions and assessing those on subsequent classifiers has successfully proved satisfactory accuracy [4]. Many experiments employ component attrition for mentoring sound and voice processing systems [5]. Nevertheless, implementations of domain adaptation in speaker recognition remain uncommon and shockingly underexplored, obviously due to a few deaths of commonly accessible platforms and well-known analyzed models [6, 7]. Representational learning and information exchange have great promise for supporting advancements in speaker recognition for various reasons [8]. Through pre-trained methods, they may well accelerate the learning and boost the signal analysis module of the audio signal Verification framework [9]. Voice search is the nearest speech analysis related to speech identification [10]. Voice control is a quasi-technology that identifies the vocabulary and grammar inside utterances [11]. Aside from enhancing its capacity to determine what is spoken, it cannot recognize or validate individuals, which are mainly indifferent in translator characteristics [12]. In reality, most audio files strive to delete person-speaking data [13]. Therefore, using recognizers alone for privacy is similar to manually entering credentials and lacking biometric identification's strength [14]. Moreover, understanding the person's words by analyzing the voice can improve the reliability of speech authenticity and minimize the authenticator [15].

Human voice relies on the interaction of the tongue and throat machine parts with physical and metabolic aspects [16]. Based on many factors, speech can be utilized as a biometric method [17]. The distinctiveness of a participant's voice is determined by the acoustic structure of a phrase, which consists of varying intensities. The multimodal recognition system combines technological innovation with biomedical studies [18], and the rapid, low-cost, seamless, and private speech biometric device attracted scientists' and technologists' interest. This comprehensive biometrics is distinguished by its rigidity, precision, and insensitivity to climatic factors [5]. Voice Recognition is a multivariate regression, similar to those other demographics technologies, with a collection of techniques that assess thousands of biological and physiological speech features [19]. The confluence of those same characteristics produces a multifactor authentication, which is a participant's distinctive voice pattern [20]. Moreover, the speech authentication algorithm analyzes the author's statement in a person's voice. It generates a credibility rating indicating whether or not the phrase and specific performance correspond to the same agent [21]. Speech biometrics is primarily programmed to automatically protect data access or resources in customer service [22]. Henceforth, a method for speaker recognition is predicated on a speech print supplied by the user during enrollment [23].

*Corresponding author.
Email address: pmghate_entc@jspmrscoe.edu.in; pravinmarotrao22@gmail.com
doi: 10.14456/easr.2023.55

A vocal trace is stored in the memory of the speaker-independent system's database. Afterwards, when the client contacts to seek admission, the method evaluates the speaker's speech in the document print [24]. Moreover, Voice recognition is distinct through other sensor authentication that identifies physical and behavioral characteristics. Organizations worldwide widely embrace these devices, including banking firms, broadband service suppliers, and other businesses that demand customer identification validation within contact center lines. Recently, speech-analysis models like large-scale deep networks [25], emotion-based signal analysis [26], etc., were implemented for this speech verification objective. But suitable verification exactness is needs found. So, speech verification for the digital application is considered the critical objective in the present work. For that, a novel CbRSI was implemented, which functioned with the support of a deep recursive network and Chimp optimal features. The critical process steps of this present work are defined as follows,

- The audio signal database was considered for this test evaluation process and imported into the system.
- Moreover, I built a novel CbRSI framework with the needed functional and speech features.
- Primarily, the noise features in the imported audio signal were analyzed and eliminated in the hidden layer using filtering parameters.
- Henceforth, the filtered audio signals were given to the classification phase as the optimal Chimp solution selected the input and the top features.
- Finally, the speech features were analyzed and matched with the saved audio features to determine the original users.
- Subsequently, the robustness parameters were restrained regarding the accuracy, sensitivity, precision, error rate and F-score.

The Speech verification framework is arranged as the recent works are detailed in the second section. The issues in the Speech verification framework are exposed third section. The novel solution for tackling those issues is given in the fourth section. The outcome gained by the execution process is figured in the fifth section, and the research discussion is concluded in section 6.

2. Related works

A few recent studies of Speech processing are determined as follows,

The large-scale features-based speech analysis framework was implemented by Chen et al. [25]. Here, the key intent of this large-scale feature is to filter the noise level of the audio input in the maximum possible range. The supervised intelligent system is utilized for training the audio database. Later, the detection features module was activated to find the noise level of the imported audio signal. Also, speech identification is executed with the help of the masking features. Hence, it visualized high filtering records, but the user verification needed to improve.

Mustaqeem and Kwon [26] has introduced a speech analysis system based on the present emotion estimation in the trained audio database. A multi-learning-based neural system is utilized for training and predicting the emotional features in the imported audio signals. Also used the sequence neural networks to analyze the speech signal continuously, and the convolutional networks extracted the features. Hence, the verification module provided a high exact outcome. However, the multiple processing models make this system complex in design.

Zhao et al. [27] developed a spatial convolutional neural system for smoothening the trained audio input. Hence, to smooth the trained audio input, the frequency and noise range of each audio signal were predicted then the desired quantity of the smoothening score was fixed. Filtered the trained audio signals based on the set smoothening frequency range; however, poor speech verification function because the smoothening role needed more features and time.

Exploring semantics could considerably increase the system throughput of semantics interactions. Hence, Weng and Qin [28] have attempted to recover received voice signals in semantics transceivers, thereby minimizing errors at the semantics phase instead of to the byte or symbols layer. So, Squeeze introduced deep networks to analyze the semantics info and identify the actual user of the particular applications. But, it needed more memory space to process the Squeeze deep networks.

Assessing a speaker's emotional situation is tricky for intelligent learning mechanisms, which have employed a crucial role in speech analysis. Speech processing is an essential topic in numerous practical uses, including human behavior evaluation and interpersonal behavior. So, Mustaqeem et al. [29] have executed the clustered speech analysis framework to verify real users in digital applications. Hence, it recorded an average speech verification score and high emotion recognition results.

Mukhamadiyev et al. [30] created a unique neural networks-based language model for speech recognition. Here the language model is trained with the Uzbek language includes 50 million phrases and 80 million words. It is created to overcome the common source language model problems. The recognition accuracy is increased significantly and solves issues in other computation linguistics. However, the classification accuracy for the word-based language model is low, and the network output unit is small for the character-based process.

The comparisons of the discussed related work are elaborated in the table format with their advantages and limitations in Table 1.

Table 1 Comparison of related works

Author	Method	Advantages	Limitations
Chen et al. [25]	Convolutional and Transformer Model (CTM)	The masking features improved the speech identification process and provided high filtering.	The user verification process needs to be enhanced.
Mustaqeem and Kwon [26]	Dilated convolutional neural system (DCNS)	The verification phase improves the accuracy.	The system is complex due to the multiple processing.
Zhao et al. [27]	Spatial convolutional neural system (SCNS)	Training and testing duration is less, suitable for real-time speech process.	The smoothening role needed more features, and its execution time was longer.
Weng and Qin [28]	Squeeze deep network (SDN)	Errors are minimized in the semantic analysis.	Need more memory space.
Mustaqeem et al. [29]	Redial-based function network (RbFN)	Recorded high-emotion recognition results	The speech verification score is average.
Mukhamadiyev et al. [30]	Neural networks	Robust for large dataset	low classification accuracy

In the past neural network-based speech processing systems, the execution time was more significant, and because of the multiprocessing, the system complexity was increased. Also, it needed more memory space. Further, due to the increased complexity, the recognition accuracy is reduced. To address these issues is developed an efficient, optimized model. Here selecting critical features based on the Chimp function reduces the system complexity in signal matching and increases the prediction accuracy. Hence, it gives the best outcome for recognizing unauthenticated signals in the speech verification system.

3. Speech verification system with problem

Suppose noises or even other design flaws mask speech data. In that case, an enlisted individual may need to submit over three or five utterances and a claim may be required to include one or more phrases for verification [31]. Conversations are necessary to request responses differently and switch to the alternative method if utterances persist in failing quality tests. Furthermore, the Customers can become annoyed or enraged by numerous calls for specimens; therefore, there needs to be a restriction on the audio signal verification counts; issues are detailed in Figure 1.

They implemented several intelligent techniques for speech signal identification based on the neural system and bio-inspired models. However, in some cases, those models were filed to identify the authenticated speech signal. It is because of noise features in the trained input audio signal. This issue is the primary concern for this speech verification system. Hence, these issues were motivated for implementing the optimized deep neural-based speech verification framework for identifying real users.

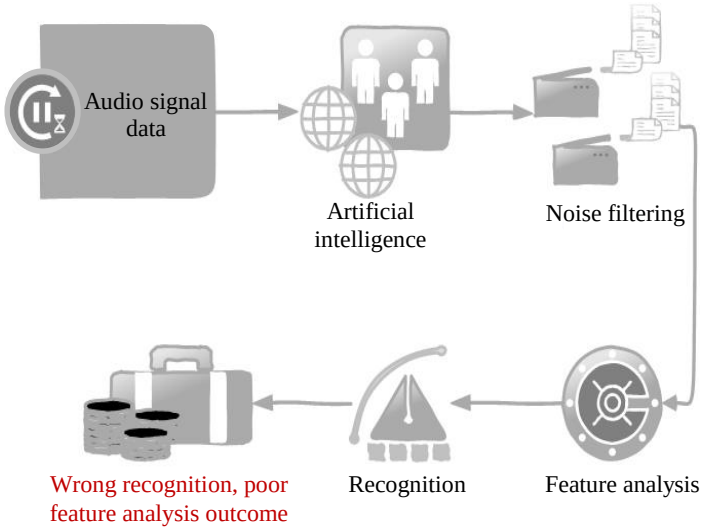

Figure 1 Speech verification system with issues

4. Proposed methodology

A novel CbRSI was executed in this research study to verify the user's speech signal by matching the saved features with the tested ones. The audio signal data is considered for this test evaluation function.

Initially, the audio data was filtered in the hidden layer of the novel CbRSI and imported into the classification phase; after that, the feature analysis and signal identification functions were activated. The proposed CbRSI architecture is determined in Figure 2. Finally, the robustness and the existing models' key performance parameters are estimated.

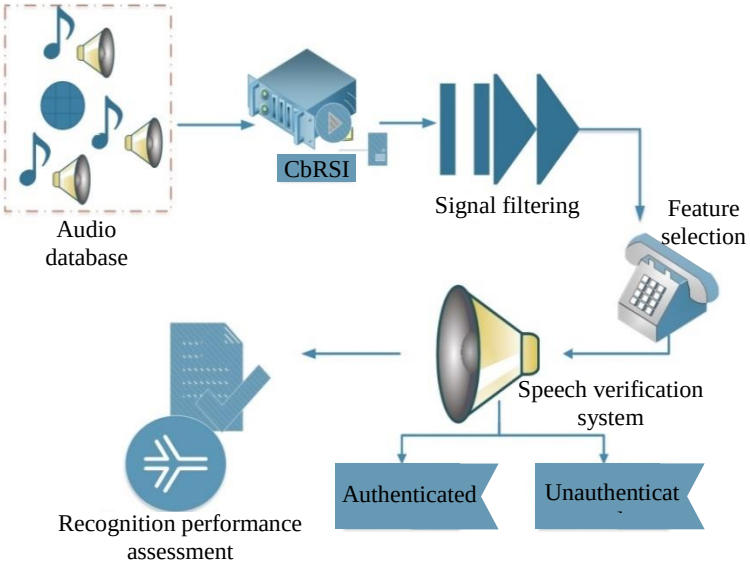

Figure 2 Proposed CbRSI Architecture

4.1 Function of CbRSI

The introduced speech prediction framework is processed based on the Chimp optimal [32] function and the recursive neural features [33]. The operational processes executed to meet the research objectives are noise features filtering, finding the required parts, and matching speech features to find the authenticated users.

$$F(A) = A\{1,2,3,4,5\ldots n\} \quad (1)$$

Here, the A defines the speech audio database and the audio signal database process is described $F(A)$. Moreover, $1,2,3,4,5\ldots n$ it represents the n count of speech signal data. Hence, the data importing process is exposed in Eqn. (1).

The novel CbRSI layers model is exposed in Figure 3. It contains five processing phases: data importing, noise features analysis, and removal in the hidden phase; feature selection and verification are made in the classification phase. Moreover, the classification phase is tuned by the optimal Chimp solution.

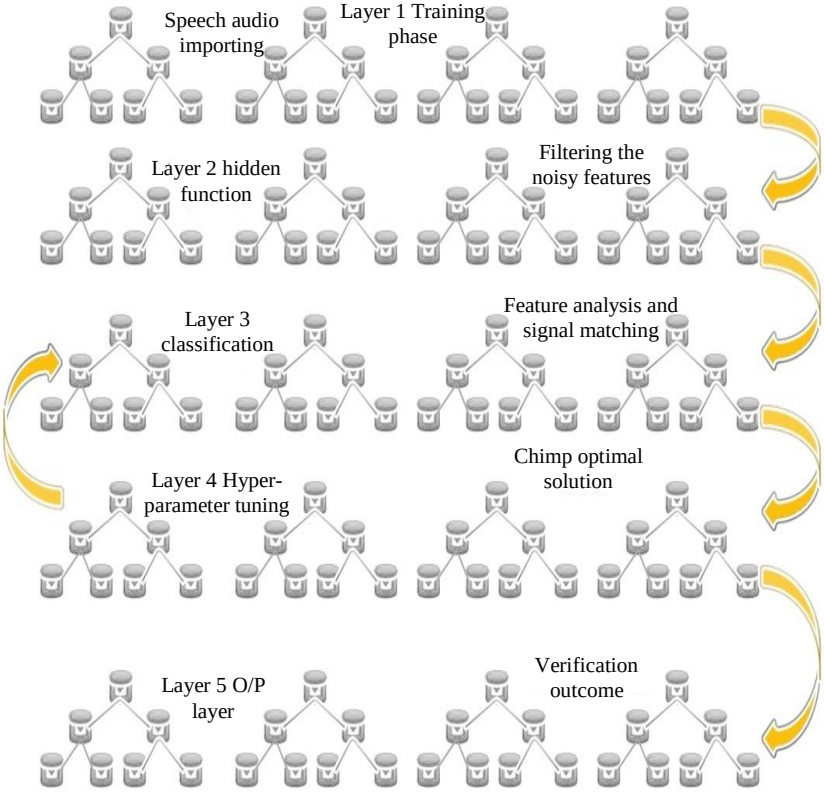

Figure 3 Layers of CbRSI

4.1.1 Preprocessing

The noisy raw audio signals might degrade the performance of the speech verification framework. Hence, it performed the filtering functions to control the degradation range and eliminate the noisy features in the imported audio signal.

$$V = |A(f - c)| \quad (2)$$

The noise filtering variable is exposed as V representing the noise features as c; the formulation specified in Eqn. (2) the raucous signal sorts were eliminated during the noise removal function and started this filtering function in the hidden phase of the novel CbRSI network. Hence, the filtering process is activated in every status for both training and testing.

4.1.2 Feature analysis

They filled the raw signal data with more features; from those features, they must select the required components for processing the speech verification strategy.

$$j = 2b(A) - A \quad (3)$$

For the verification system, the binary output is the most critical parameter for analyzing the testing signal for authentication purposes. In addition, the essential significance of the feature selection module is to reduce the complexity range of the speaker verification system. If the required features are not selected, the identification process must be longer to find the needed audio features and match them with the saved features. Considering these drawbacks, feature selection is a significant step for all intelligent model-based prediction applications.

Algorithm 1 CbRSI

```
Start ()
{
    int A = 1,2……n
    // training audio signal data
    Preprocessing ()
    {
        int V, f, c;
        // initializing the filtering variable for neglecting the noisy features
        filtering(A) =|A – c|
        // Noisy features are removed by performing the filtering process
    }
    Feature analysis ()
    {
        int j, b;
        // feature selection variables are initialized
        select ⟶ 2b(A)
        // selecting the required features
    }
    Speech verification ()
    {
        if (j(i) = j(l))
        {
        Authenticated
        }
        Else (unauthenticated)
    }
}
Stop ()
```

4.1.3 Speech verification

Once the filtering and the feature selection function for the trained audio input are completed, the speech verification activated operational modules to find the authorized and third parties. Here, the speech verification variable is described as P. Moreover, the trained audio data features are exposed $j(i)$ and the testing audio signal features are defined as $j(l)$.

$$P = \begin{Bmatrix} if\left(j(i) = j(l)\right) & Authenticated \\ if\left(j(i) \neq j(l)\right) & Unauthenticated \end{Bmatrix} \tag{4}$$

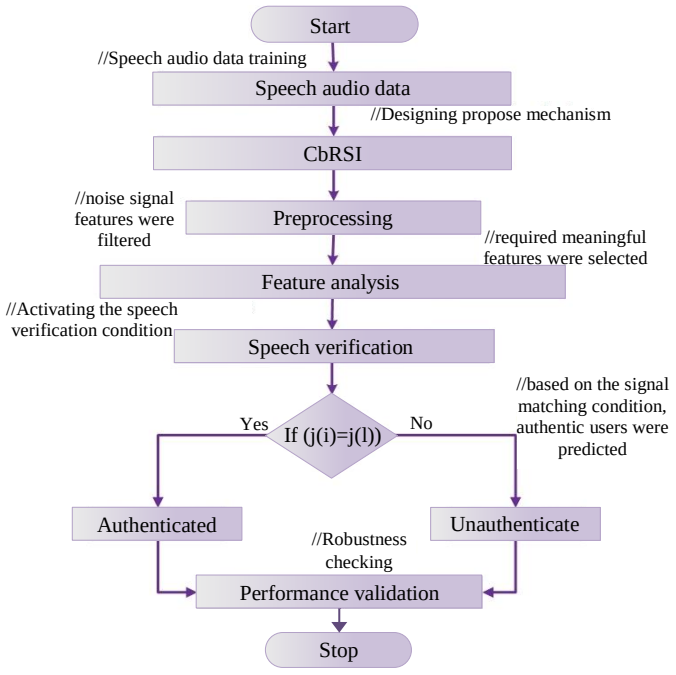

Figure 4 Flow of CbRSI

Here, the users are authorized by matching the spectrogram of the tested audio signal with the saved audio spectrogram features. Hence, the matching evaluation condition is defined in Eqn. (4), which is adapted from the chimp fitness solution.

The defined mathematical formulations of each process are arranged algorithmically, defined in algorithm 1, and the operation flow steps in order wise are defined in Figure 4. Several deep networks and bio-inspired models are already defined for the mathematical and problem-solving studies. Still, the reason for attaining the recursive networks is their flexibility passed on the different trained input parameters, resulting in a high exactness score. In addition, the chimp optimal is selected based on their unique hunting behavior by tracing the prey. Here, that intelligent function is utilized to track the required signal features.

5. Results and discussion

The speech analyzing system is validated in the MATLAB programming environment, and the robustness was valued by processing the audio signal database. It included 2530 authorized audio signals and 1005 unauthorized audio. The database was validated in the ratio of 75% training and testing 25%, which contains both authorized and unauthenticated speech signals. The collected datasets are initially preprocessed to remove the signal noises and converted into binary format. Furthermore, the required features are extracted by the four prey-tracing functions of the chimp. The extracted features are updated at the prediction phase, and signal matching is performed. Based on the matching parts, the signals are classified into authenticated and unauthenticated at the speech verification phase. The Execution constraints required for the CbRSI processing are defined in Table 2.

Table 2 Execution constraints

Parameter description	
Input	Speech audio database
Total audio signal	3535
Software platform	MATLAB R2021a
Database format	WAV audio
Running environment	Windows 10
Processor	1 gigahertz (GHz) or faster processor or System on a Chip (SoC)
RAM	1 gigabyte (GB) for 32-bit or 2 GB for 64-bit
Hard drive space	16 GB for 32-bit OS 32 GB for 64-bit OS
Application	Authenticated speech signal verification
Size	2D magnitude spectrogram
Training network	Recursive neural network
Algorithm	Chimp optimization

5.1 Case study

To check the operation performance of the novel CbRSI, I used the working function of the book. CbRSI is evaluated for both different classes that are authenticated and unauthenticated user classes and the process is explained with the signal output samples.

Here, the saved audio spectrogram matches the tested preprocessing audio signal. Hence, it is recognized as the authenticated user. In addition, the input amplitude of the sampled audio signal is exposed in Figure 5 a). Here, the amplitude range shows both ranges are stable while raising and lowering.

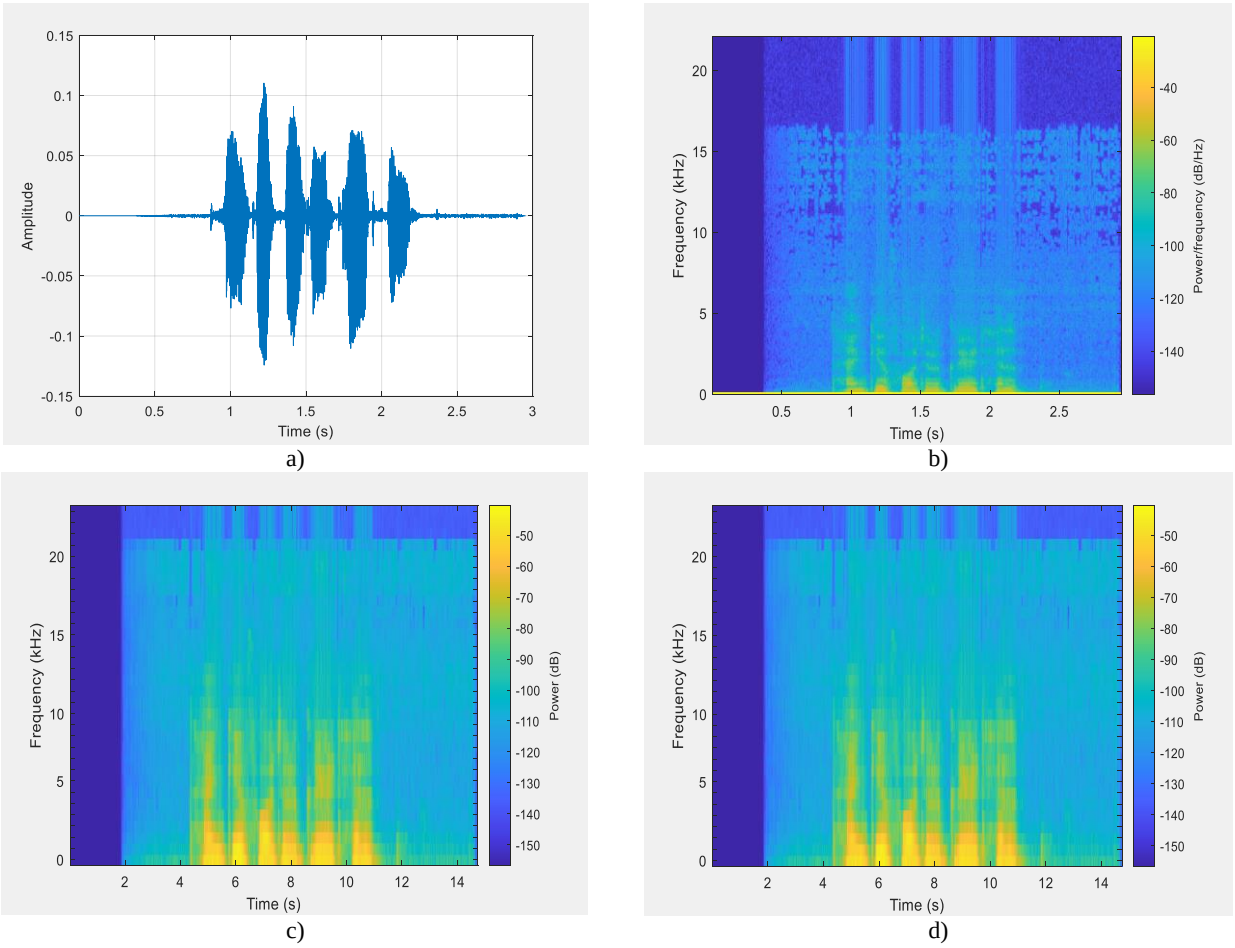

Figure 5 Authenticated signal: a) input, b) spectrogram of the input signal, c) filtered spectrogram, d) saved spectrogram

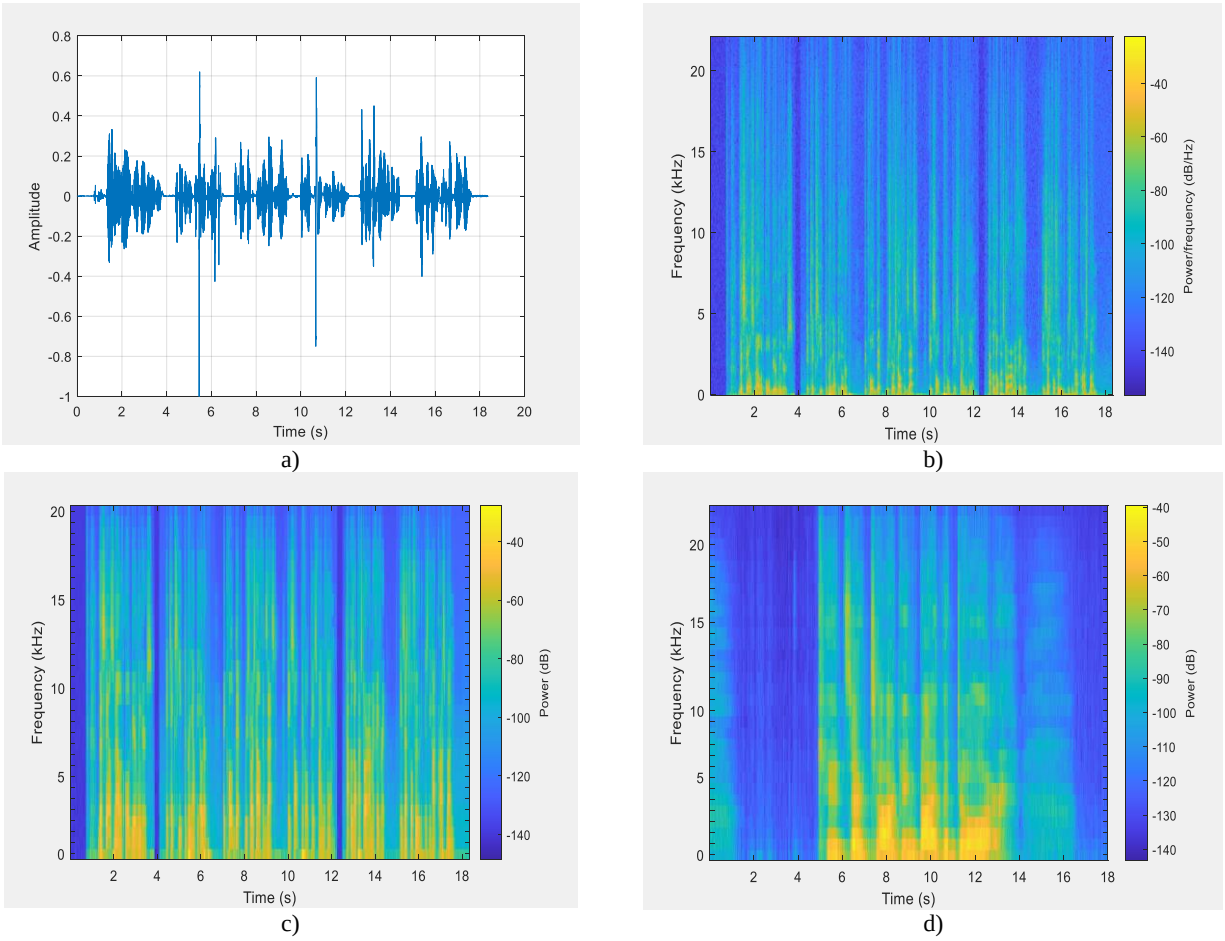

Figure 6 Unauthenticated speech signal: a) input, b) input audio signal's spectrogram, c) preprocessed spectrogram, d) saved spectrogram

Here, the saved audio spectrogram in the memory of the novel CbRSI does not match the tested signal, displayed in Figure 6, c) and Figure 6 d). Henceforth, it is recognized as an unauthenticated signal.

5.2 Performance assessment

The novel CbRSI robustness score is determined by evaluating the critical parameters like F-score, accuracy, error, recall and precision. Henceforth, the past studies considered for this comparative measure are One-shot Neural Networks (ONN) [34], Statistics Convolutional Neural systems (SCNS) [35], and Recursive Network Systems (RNS) [33].

5.2.1 Accuracy

The exactness score of the speech signal identification is evaluated through Eqn. (5). Earned the accuracy score with the correct speech recognition average from the total tested signal.

$$Accuracy = \frac{Correct\ recognition}{Total\ audio\ input} \tag{5}$$

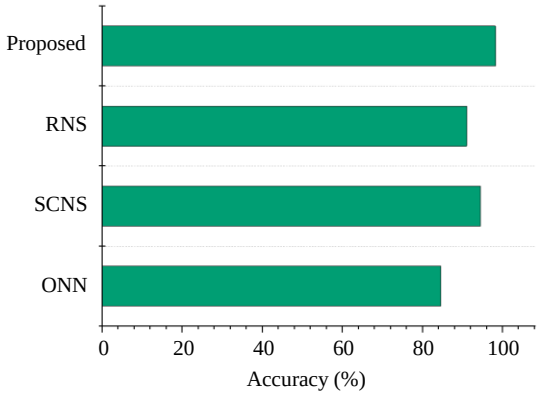

Figure 7 Accuracy evaluation

The exactness value of the model ONN is 84.6%, SCNS yielded 94.5%, and the RNS scored 91%. Considering these mechanisms, the novel CbRSI attained the full exactness measure of 98.2%. These assessments are exposed in Figure 7.

5.2.2 Precision evaluation

The average accuracy of the positive speech recognition for both true and false classes is measured by the precision evaluation metrics detailed in Eqn. (6).

$$Precision = \frac{True\_positives}{False\_positives + True\_positives} \quad (6)$$

The precision score earned by the RNS is 98%, ONN reported the maximum precision range as 91%, and the SCNS yielded 94.3% precision validation. Besides, the novel CbRSI gained 98% precision, detailed in Figure 8. It is the most comprehensive precision score of the compared model.

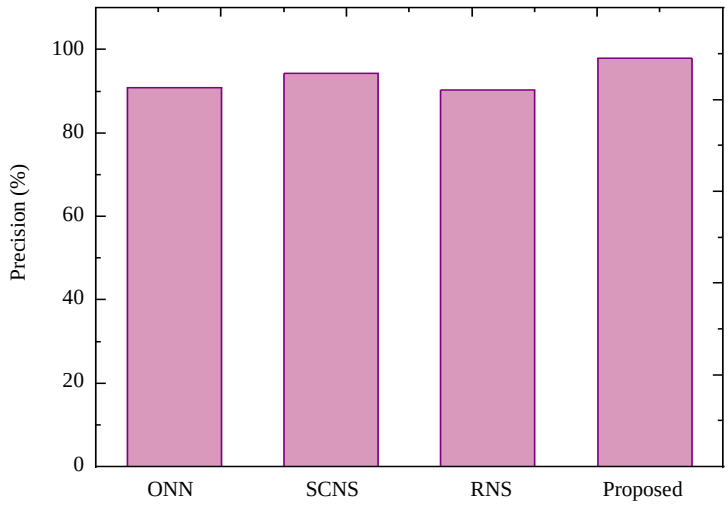

Figure 8 Precision measurement

5.2.3 Recall validation

The trained audio data includes dual classes that are authorized and unauthenticated users; based on these class features, the specification of the true and false class samples was determined. The recall performance parameter is evaluated to validate the negative identification outcome of the actual class samples, and the formulation of recall is detailed in Eqn. (7).

$$Recall = \frac{True\_Negatives}{False\_Negatives + True\_positives} \quad (7)$$

The recall rate of the SCNS approach is 96.04%, the RNS mechanism earned 90% recall, and the model ONN employed a recall score of 85%. Considering all past studies, the novel CbRSI was measured at the finest recall range of 98.2%. Moreover, the recall estimation is exposed in Figure 9.

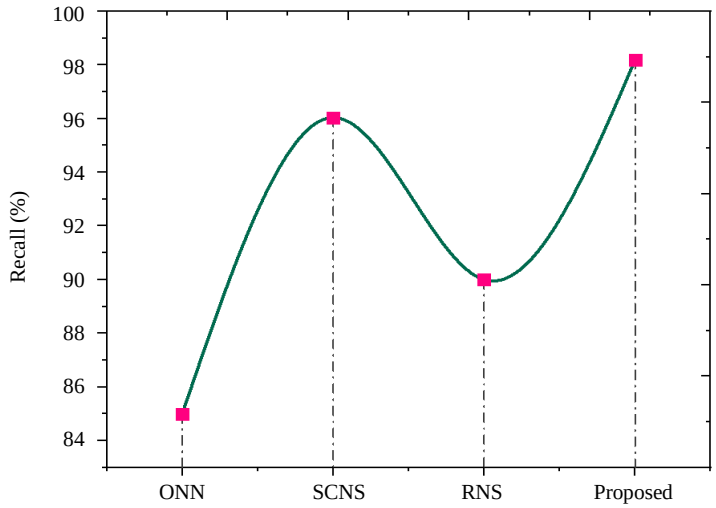

Figure 9 Recall Evaluation

5.2.4 Evaluation of F-measure

The exactness score for the prediction function is validated using different parameters based on trained applications. Henceforth, to determine the average recognition robustness score, the f-value is evaluated. Eqn. (8) Estimate the mean speech identification exactness range. If the recall and precision occurrence are good, then the F-measure yielded the highest performance.

$$F\_value = \frac{2 \times Recall \times Precision}{Precision + Recall} \tag{8}$$

The yielded F-value by the RNS is 90.5%, SCNS has attained 94.6% F-measure and the mechanism ONN defined 83% F-value. Besides, the novel CbRSI reported the outstanding F-value as 98.1%; performance is visualized in Figure 10.

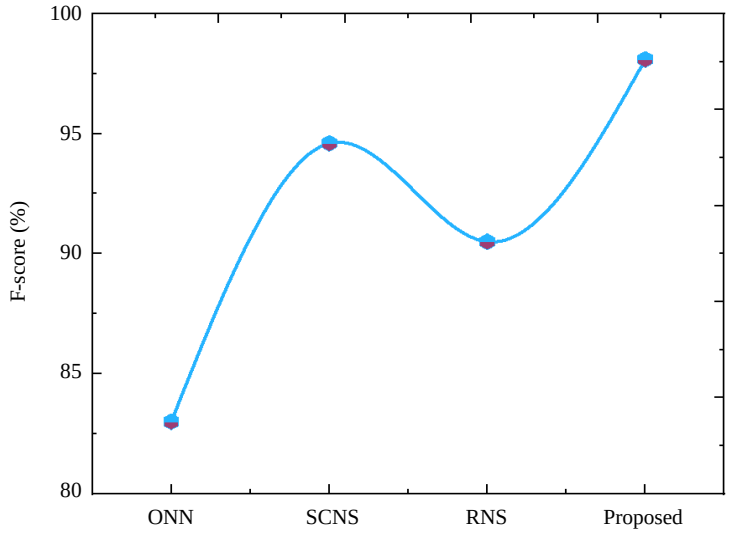

Figure 10 F-value assessment

5.2.5 Assessment of error rate

For any mechanism, justification is essential to estimate the need for the presented novel solution for the particular applications. In addition, the fall rate of the present method can help the following researchers to find a new efficient solution for the specific domain. Hence, the novel CbRSI is justified through the fall rate of the designed system. Henceforth, the error measure is formulated by Eqn. (9).

$$Error = \frac{Wrong\ Prediction}{Total\ Prediction} \tag{9}$$

The lowest error score recorded for the novel CbRSI is 1.8%, quite the lowest error score compared to past studies, such as ONN, SCNS and RNS. Here, the RNS reported a 9% error score, ONN yielded a 15.4% error score, and the model SCNS gained a 5.6% error value, as defined in Figure 11. The Comparison details are described in Table 3.

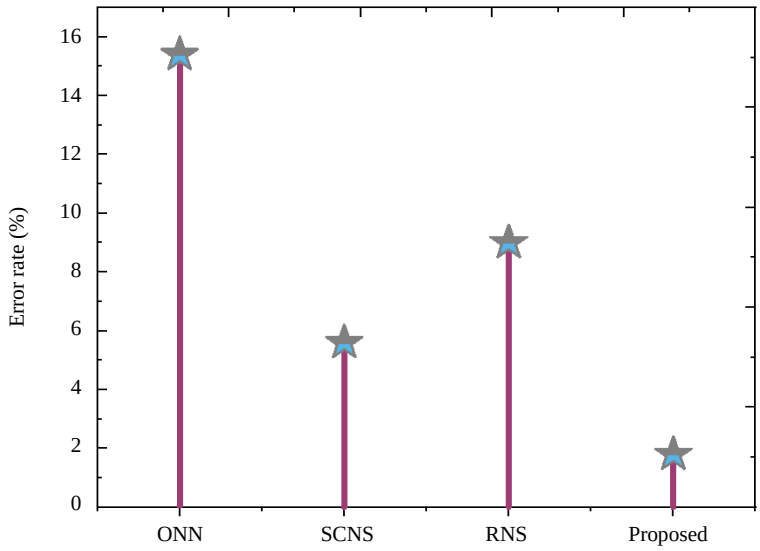

Figure 11 Error rate estimation

Table 3 Comparison details

Performance statistics					
Methods	**F-score (%)**	**Precision (%)**	**Accuracy (%)**	**Error rate (%)**	**Recall (%)**
ONN	83	91	84.6	15.4	85
SCNS	94.6	94.3	94.5	5.6	96.04
RNS	90.5	90.3	91	9	90
Proposed	98.1	98	98.2	1.8	98.2

The novel CbRSI scored the finest outcome from the estimated performance parameters in all robustness assessments. Henceforth, the book CbRSI is suitable for the speech verification domain for finding the real user among unauthorized users. Additionally, the performance of the related speech signal processing models, such as CTM, DCNS, SCNS, SDN and RbFN, were compared with the proposed speech verification model.

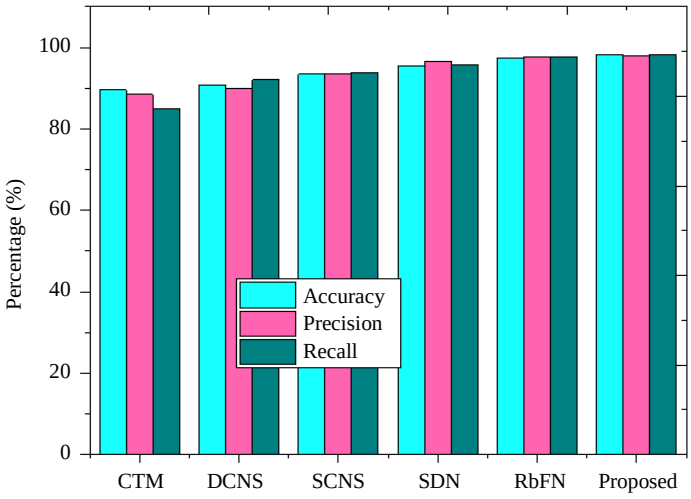

Figure 12 Accuracy, recall and precision evaluation

The precision, recall and accuracy comparison is given in Figure 12. The accuracy of the related models such as CTM, DCNS, SCNS, SDN and RbFN is 89.5%, 90.8%, 93.4%, 95.34% and 97.5%. The precision values are 88.4%, 89.97%, 93.65%, 96.5% and 97.56%. Also, the recall measures are validated as 85%, 92%, 93.8%, 95.8% and 97.54%. From the comparison with the state of art models, the proposed signal processing method gained higher values and verified the efficiency of the unauthenticated speech identification.

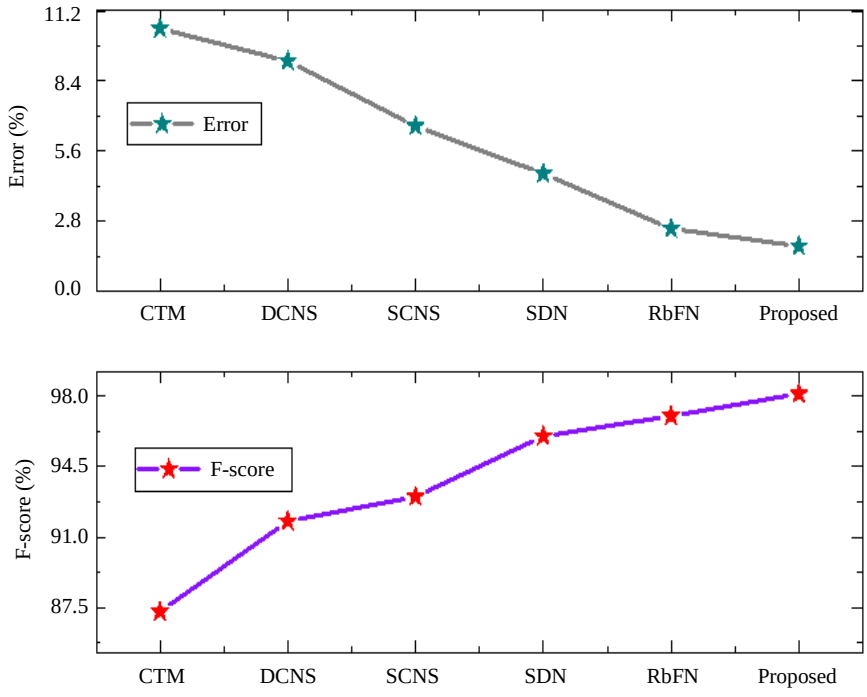

Figure 13 F-score and error comparison

Also, the f-score and error rate comparison is shown in a graphical representation in Figure 13. Here, the f-score of the prevailing models, such as CTM, DCNS, SCNS, SDN and RbFN, is 87.3%, 91.78%, 93%, 96% and 97%. The attained error values of these prevailing models are 10.5%, 9.2%, 6.6%, 4.7% and 2.5%. At the same time, the proposed CbRSI validated the f-score and error rate as 98.1% and 1.8%. The comparison clearly shows that the proposed model attained a higher f-score, and reduced the error rate. The higher value of the f-score indicates better recall and precision results and efficient system performance.

To verify the time efficiency, calculated in seconds the run time of the proposed CbRSI model and contrasted it to the other recent related models in Figure 14. Here, the model CTM scored the execution time as 2.4s; DCNS achieved the results in 2.278s; SCNS completed the process in 1.95, and SDN and RbFN validated the run time as 1.782s and 1.59s. Compared to other models, the designed method achieved less run time, measured as 1.32. The selection of the required features through the Chimp function reduced the complexity of the signal-matching module. Hence the run time in the designed module is reduced.

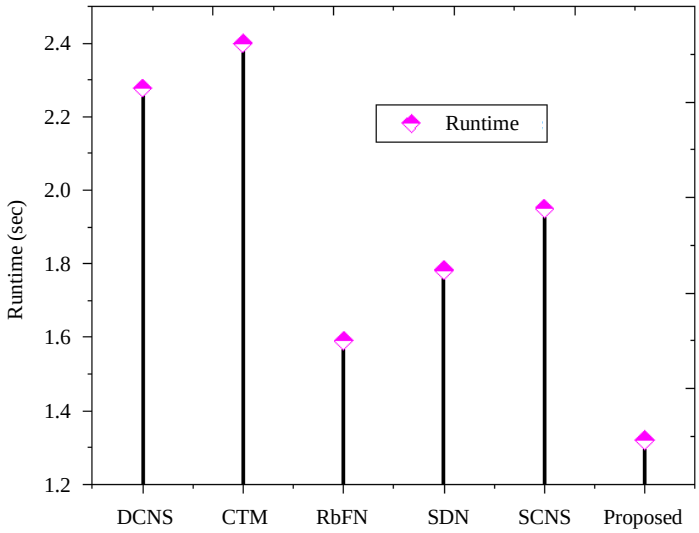

Figure 14 Run time assessment

Table 4 State-of-the-art design performance comparison

Method	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)	Error (%)	Runtime (s)
CTM	89.5	88.4	85	87.3	10.5	2.4
DCNS	90.8	89.97	92	91.78	9.2	2.278
SCNS	93.4	93.65	93.8	93	6.6	1.95
SDN	95.34	96.5	95.8	96	4.7	1.782
RbFN	97.5	97.56	97.54	97	2.5	1.59
Proposed	98.2	98	98.2	98.1	1.8	1.32

The overall comparison of the state of the approaches is arranged in Table 4. The enhanced metrics value showed the better performance of the CbRSI in the exact prediction of the authenticated and the unauthenticated signal for the speech verification system. Also, in the proposed method, the unauthenticated movement is predicted within the short duration that is verified by the lower run time value.

5.3 Discussion and limitation

In this article, a novel CbRSI design was developed to verify the authenticated and unauthenticated signals. The model follows preprocessing, feature extraction, matching and prediction process sequentially. This system has provided a better outcome in identifying the unauthorized, verified by the performance analysis. The model attained higher accuracy than other models and achieved better classification results. Thus the system is efficient for speech verification to prevent third-party access. Using the recursive function gives better learning ability to the system, and the optimization function increases the overall accuracy and reduces the computation duration. It is also suitable for processing long-length inputs. However, the tree structure of each piece of information needs to be identified at the learning stage. The utilized optimization function often falls into local optima and minimum for complex inputs. The selection of a better processing network with an efficient optimization model in future can fill these gaps.

6. Conclusion

The novel speech verification system CbRSI was implemented for this present research study. The speech audio database was considered for estimating the robustness of the novel CbRSI framework. The function process, like filtering feature analyzing and speech features matching, is performed with optimal chimp fitness. Henceforth, the executed approach earned the widest audio signal identification exactness at 98.2%; compared to the past studies, it improved by 4% of the recognition exactness. In addition, the score least wrong speech recognition rate by the novel CbRSI mechanism is 1.8%, which also determines a 4% error reduction than the past associated models. Also, the metrics recall, precision, and f-score have defined a 4% improvement over the compared associated frameworks. However, the drawback behind the implemented model is the robustness of analyzing against spoofing attacks in the speech processing system. In future, incorporating the spoofing attack features and validating the performance of the implemented model will reveal the reliability score.

7. References

[1] Islam MA, Sakib AN. Bangla dataset and MMFCC in text-dependent speaker identification. Eng Appl Sci Res. 2019;46(1):56-63.

[2] Myint LMM, Warisarn C, Busyatras W, Kovintavewat P. Single-Track equalization method with TMR correction system based on cross correlation functions for a patterned media recording system. Eng Appl Sci Res. 2017;44(1):16-9.

[3] Hamcumpai S, Bureerat S, Eua-Anant N. Comparison of signal processing techniques for fault detection in helical spur gears. KKU Eng J. 2007;34(1):59-72.

[4] Stafylakis T, Mošner L, Kakouros S, Plchot O, Burget L, Ćernocký J. Extracting speaker and emotion information from self-supervised speech models via channel-wise correlations. 2022 IEEE Spoken Language Technology Workshop (SLT); 2023 Jan 9-12; Doha, Qatar. USA: IEEE; 2023. p. 1136-43.

[5] Korkmaz Y, Boyacı A. Hybrid voice activity detection system based on LSTM and auditory speech features. Biomed Signal Process Control. 2023;80:104408.

[6] Abdusalomov AB, Safarov F, Rakhimov M, Turaev B, Whangbo TK. Improved feature parameter extraction from speech signals using machine learning algorithm. Sensors. 2022;22(21):8122.

[7] Ren D, Srivastava G. A novel natural language processing model in mobile communication networks. Mobile Netw Appl. 2022;27:2575-84.

[8] Kwon H, Nam SH. Audio adversarial detection through classification score on speech recognition systems. Comput Secur. 2023;126:103061.

[9] Zheng WZ, Han JY, Cheng HL, Chu WC, Chen KC, Lai YH. Comparing the performance of classic voice-driven assistive systems for dysarthric speech. Biomed Signal Process Control. 2023;81:104447.

[10] Yang Y, Zhang H, Cai Z, Shi Y, Li M, Zhang D, et al. Electrolaryngeal speech enhancement based on a two stage framework with bottleneck feature refinement and voice conversion. Biomed Signal Process Control. 2023;80:104279.

[11] Madhu H, Satapara S, Modha S, Mandl T, Majumder P. Detecting offensive speech in conversational code-mixed dialogue on social media: a contextual dataset and benchmark experiments. Expert Syst Appl. 2023;215:119342.

[12] Meng W, Yolwas N. A study of speech recognition for Kazakh based on unsupervised pre-training. Sensors. 2023;23(2):870.

[13] de Lope J, Graña M. An ongoing review of speech emotion recognition. Neurocomputing. 2023;528:1-11.

[14] Yang CC, Chang TS. A 1.6-mW sparse deep learning accelerator for speech separation. IEEE Trans Very Large Scale Integr (VLSI) Syst. 2023;31(3):310-9.

[15] Dowerah S, Serizel R, Jouvet D, Mohammadamini M, Matrouf D. Joint optimization of diffusion probabilistic-based multichannel speech enhancement with far-field speaker verification. 2022 IEEE Spoken Language Technology Workshop (SLT); 2023 Jan 9-12; Doha, Qatar. USA: IEEE; 2023. p. 428-35.

[16] Lin W, Mak MW. Robust speaker verification using deep weight space ensemble. IEEE/ACM Trans Audio Speech Lang Process. 2023;31:802-12.

[17] Abbasi W. Privacy-Preserving speaker verification and speech recognition. In: Saracino A, Mori P, editors. Emerging Technologies for Authorization and Authentication. Lecture Notes in Computer Science, vol. 13782. Cham: Springer; 2023. p. 102-19.

[18] Cai Z, Yang Y, Li M. Cross-lingual multi-speaker speech synthesis with limited bilingual training data. Comput Speech Lang. 2023;77:101427.

[19] Mingote V, Miguel A, Ortega A, Lleida E. Class token and knowledge distillation for multi-head self-attention speaker verification systems. Digit Signal Process. 2023;133:103859.

[20] Li J. A comparative study of different filters for speech signals. International Conference on Intelligent Systems, Communications, and Computer Networks (ISCCN 2022); 2022 Jun 17-19; Chengdu, China. Washington: SPIE; 2022. p. 12332-6.

[21] Chen Z, Yoshioka T, Lu L, Zhou T, Meng Z, Luo Y, et al. Continuous speech separation: dataset and analysis. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020 May 4-8; Barcelona, Spain. USA: IEEE; 2020. p. 7284-8.

[22] Xia Y, Braun S, Reddy CKA, Dubey H, Cutler R, Tashev I. Weighted speech distortion losses for neural-network-based real-time speech enhancement. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020 May 4-8; Barcelona, Spain. USA: IEEE; 2020. p. 871-5.

[23] Maciejewski M, Wichern G, McQuinn E, Roux JL. WHAMR!: Noisy and reverberant single-channel speech separation. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020 May 4-8; Barcelona, Spain. USA: IEEE; 2020. p. 696-700.

[24] Michelsanti D, Tan ZH, Zhang SX, Xu Y, Yu M, Yu D, et al. An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Trans Audio Speech Lang Process. 2021;29:1368-96.

[25] Chen S, Wang C, Chen Z, Wu Y, Liu S, Chen Z, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE J Sel Top Signal Process. 2022;16(6):1505-18.

[26] Mustaqeem, Kwon S. MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. Expert Syst Appl. 2021;167:114177.

[27] Zhao Y, Wang DL, Xu B, Zhang T. Monaural speech dereverberation using temporal convolutional networks with self-attention. IEEE/ACM Trans Audio Speech Lang Process. 2020;28:1598-607.

[28] Weng Z, Qin Z. Semantic communication systems for speech transmission. IEEE J Sel Areas Commun. 2021;39(8):2434-44.

[29] Mustaqeem, Sajjad M, Kwon S. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE Access. 2020;8:79861-75.

[30] Mukhamadiyev A, Mukhiddinov M, Khujayarov I, Ochilov M, Cho J. Development of language models for continuous Uzbek speech recognition system. Sensors. 2023;23(3):1145.

[31] Venkateswarlu SC, Kumar NU, Veeraswamy D, Vijay V. Speech intelligibility quality in telugu speech patterns using a wavelet-based hybrid threshold transform method. In: Reddy VS, Prasad VK, Mallikarjuna Rao DN, Satapathy SC, editors. Intelligent Systems and Sustainable Computing. Smart Innovation, Systems and Technologies, vol. 289. Springer: Singapore; 2022. p. 449-62.

[32] Khishe M, Nezhadshahbodaghi M, Mosavi MR, Martín D. A weighted chimp optimization algorithm. IEEE Access. 2021;9:158508-39.
[33] Wan S, Yeh ML, Ma HL, Chou TY. The robust study of deep learning recursive neural network for predicting of turbidity of water. Water. 2022;14(5):761.
[34] Saxena N, Varshney D. Smart home security solutions using facial authentication and speaker recognition through artificial neural networks. Int J Cogn Comput Eng. 2021;2:154-64.
[35] Nainan S, Kulkarni V. Enhancement in speaker recognition for optimized speech features using GMM, SVM and 1-D CNN. Int J Speech Technol. 2021;24:809-22.