

Discriminative deep learning based hybrid spectro-temporal features for synthetic voice spoofing detection

Pranita Niraj Palsapure, Rajeswari, Sandeep Kumar Kempegowda, Kumbhar Trupti Ravikumar

Department of Electronics and Communication Engineering, Acharya Institute of Technology, Visvesvaraya Technological University, Belagavi, India

Article Info

Article history:

Received Oct 16, 2023

Revised Jan 11, 2024

Accepted Feb 21, 2024

Keywords:

Automatic speaker verification

Hybrid feature learning

LSTM-CNN

Spectral-temporal feature

Spoofing attack detection

ABSTRACT

Voice-based systems like speaker identification systems (SIS) and automatic speaker verification systems (ASV) are proliferating across industries such as finance and healthcare due to their utility in identity verification through unique speech pattern analysis. Despite their advancements, ASVs are susceptible to various spoofing attacks, including logical and replay attacks, posing challenges due to the sophisticated acoustic distinctions between authentic and spoofed voices. To counteract, this study proposes a robust yet computationally efficient countermeasure system, utilizing a systematic data processing pipeline coupled with a hybrid spectral-temporal learning approach. The aim is to identify effective features that optimize the model's detection accuracy and computational efficiency. The model achieved superior performance with an accuracy of 99.44% and an equal error rate (EER) of 0.014 in the logical access scenario of the ASVspoof 2019 challenge, demonstrating its enhanced accuracy and reliability in detecting spoofing attacks with minimized error margin.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Pranita Niraj Palsapure

Department of Electronics and Communication Engineering, Acharya Institute of Technology

Visvesvaraya Technological University

Belagavi 590018, Karnataka, India

Email: pranitanirajpalsapure@gmail.com

1. INTRODUCTION

In today's digital world, where personal data and security are paramount, biometric authentication delivers a unique and convenient way to identify individuals based on their physiological traits rather than traditional passwords and PINs [1], [2]. Among various biometric modalities, voice-based systems are a promising technology with several advantages compared to other biometric modalities such as fingerprint and facial recognition [3]. Voice-based authentication-based approaches are more user-friendly because they do not require physical contact from the user, thus preserving privacy and allowing for frictionless user authentication in remote environments [4]. Basically, voice-based biometric applications employ a speaker recognition system (SIS) and automatic speaker verification system (AVS). SIS are designed to determine a speaker's identity, while ASV is used to confirm that the presented speaker's identity is who he claims to be [5]. ASV systems are widely used in security-sensitive industries, such as access control, user authentication, financial services, virtual assistants, telecommunications, and healthcare [6]. By analyzing speech patterns and characteristics, ASVs can prevent unauthorized access, reduce identity fraud, and improve user experience [7]. Although ASV systems have promising benefits, they are not immune to voice spoofing attacks, in which an attacker attempts to trick the system by imitating a legitimate user's voice [8]. Replay attacks involve recording a genuine user's voice and playing it back to trick the ASV system into thinking that the attacker is the genuine

speaker [9]. Additionally, speech synthesis uses text-to-speech (TTS) technologies to create speech that mimics the genuine speaker's voice features to create realistic-sounding speech to spoof the ASV system [10]. Voice transformation attacks involve modifying an attacker's voice to resemble a real speaker's, bypassing the ASV system [11]. As technology continues to evolve, attackers will increasingly adopt more sophisticated methods to deceive ASV systems, making it challenging for anti-spoofing systems to remain effective [12]. Another challenge is handling the quality of training data sets, which are often affected by high dimensions and temporal dynamics. Furthermore, authentic, and spoofed speech acoustic features may be very similar, representing an inter-class similarity problem. With these complexities and advances in artificial intelligence technology, voice spoofing attacks seriously threaten the reliability of ASV systems [13]. Hence, robust anti-spoofing measures are essential to counter various forms of this threat. This paper introduces a computationally efficient design of an intelligent voice spoofing countermeasure system for ASV that leverages a systematic data processing pipeline, multi-level audio features modelling and the strengths of a hybrid spectral-temporal learning model to classify the genuine and spoofed voice effectively.

In a recent state of the art, significant research has been done to address the challenges posed by voice spoofing attacks. Researchers have explored many anti-spoofing techniques, including feature-based, deep-learning approaches and sophisticated models. Magazine *et al.* [14] highlighted the potential threat of deepfakes to ASV technologies. They introduced a detection approach on modulation-spectrogram features that characterizes session identity, gender, and the source of generation variation to differentiate between them. Zhang *et al.* [15] introduced a one-class learning method to detect TTS-generated spoofed voices. Their approach revolves around compacting genuine speech representations and introducing an angular margin, without data augmentation, to distinguish between genuine and spoofed voices in the embedding space. When tested on the ASVspoof 2019 dataset, their model achieved an equal error rate (EER) of 2.19%. An application of ensemble learning for voice anti-spoofing system is presented by Dua *et al.*, [16] where recurrent neural network (RNN) and convolutional neural network (CNN) were trained on Mel-frequency-cepstral-coefficients (MFCC) extracted from the ASVspoof dataset. Zhang *et al.* [17] explored a voice spoofing scenario where fake voice signals generated using TTS are embedded into legitimate utterances. The authors developed a countermeasure scheme utilizing self-pre-trained models for feature extraction. Additionally, they aimed for simultaneous utterance- and segment-level detection. Experimentally, their approach achieved EERs of 0.77% on the PartialSpoof database and 0.90% on ASVspoof 2019.

In the study of Xue *et al.* [18], an iterative knowledge distillation method is adopted for fake speech detection where a deep network as the instructor is modelled to guide multiple shallow classifiers by minimizing feature differences. Lei *et al.* [19] developed a method to detect known and unknown spoofing attacks using 1-D CNN, Siamese CNN, and Gaussian mixture model (GMM) components to capture local and global speech features. Wu *et al.* [20] introduced the feature engineering technique, which uses a transformer trained on a genuine speech from the ASVspoof 2019 logical access corpus to identify and remove spoofing artefacts. Javed *et al.* [21] developed a framework that uses co-occurrence patterns and cepstral coefficients to detect distortions and artefacts induced by different spoofing methods, providing comprehensive protection against even complex spoofing attacks. The effectiveness of hybridizing different learning models, namely CNN for feature extraction and support vector machine (SVM) for classification tasks, is studied by [22]. Kwak *et al.* [23], [24] developed new models that are more efficient and robust to unseen spoof attacks. Guo *et al.* [25] used incremental learning to improve the generalizability of spoof detection models to unseen spoof algorithms. They discuss how to enhance these models' embedding space and decision boundaries to adapt to new spoofing threats. Malik *et al.* [26] addressed the vulnerability of voice-activated services like chatbots to audio replay attacks. They introduced acoustic ternary patterns-gammatone cepstral coefficient and used a multi-class SVM classifier trained through error-correcting output codes on the optimal feature space. The model presented by Adiban *et al.* [27] utilizes one-class learning for detecting synthetic voice spoofing. An EER of 2.19 depicts moderate efficiency in minimizing false acceptance and rejection rates. Hence, it can be seen that researchers have proposed various strategies to counteract voice spoofing attacks, yet many existing methods involve complex and sophisticated models. Most current approaches predominantly focus on extracting mel-spectrogram features and often employ CNN for classification. However, exploring alternative features representing both temporal and spectral signal characteristics could potentially enhance the efficiency of detecting voice spoofing attacks on ASV systems.

2. METHOD

This section details the proposed system design and elaborates on the implementation procedure for voice spoof countermeasures for ASV systems. First, the adopted dataset is briefly described, and following this, the extraction of the essential features that contribute towards leveraging more weightage in the latent feature extraction for the proposed hybrid spectral-spatial learning model. Further, an implementation strategy was adopted for feature matrix preparation and training the hybrid learning model is discussed.

2.1. Proposed system

The research work reported in this paper is primarily focused on introducing a unique and lightweight approach to developing a sustainable voice spoof countermeasure system for real-world ASV system applications. The proposed system is designed to dynamically adapt to the speaker's voice signal deviations, regardless of the inter-class similarity between the spoof and the real audio signal. In this case, the study believes that if an attacker spoofs or mimics the audio of the real speaker, he also has some latent feature that, if identified, can help him learn the pattern of the spoofed signal. Can and be able to create a learning model for efficiently. Reduce opposing speech. Therefore, the proposed study focuses on extracting multiple features from the audio files, as different features provide different voice attributes, often in the form of spectral and temporal representations. The proposed study implements a hybrid spectral-temporal learning model by leveraging CNN and long short-term memory (LSTM) deep learning models. The CNN model is used to learn and capture the spectral data, while LSTM has been implemented to learn temporal features and capture the long-term dependencies. The proposed system design is lightweight because the study performs dimensionality reduction operations on some extracted features and focuses on implementing a highly optimized learning model with optimal learnable parameters, with longer training epochs and optimal hyperparameters selection. The schematic architecture of the proposed system is shown in Figure 1.

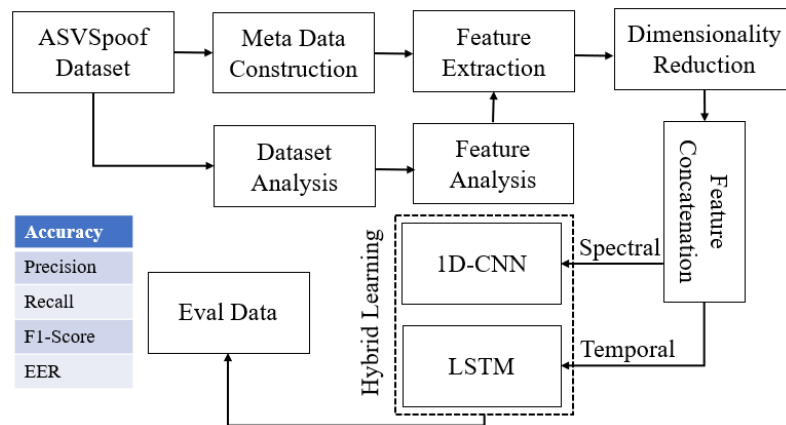


Figure 1. A proposed hybrid spectral-temporal model for spoof countermeasure

Figure 1 depicts the block-based workflow of the proposed hybrid spectral-temporal model, which comprises four core modules such as: i) exploratory data analysis, ii) meta-data construction, iii) feature extraction, iii) feature concatenation, and iv) learning spectral and temporal features for spoof countermeasure. The novelty of the proposed system is that it uses a lightweight approach that is not computationally expensive. This feature makes it more practical for real-world applications. The system is sustainable because it can adapt to the speaker's voice signal changes over time. This aspect is important because the characteristics of a speaker's voice can change due to factors such as age, illness, or environmental conditions. The system is practical for real-world ASV system applications because it can be implemented on various hardware platforms and does not require much training data.

2.2. Dataset description

The dataset considered in this study is ASVSpooF 2019 [28], a large-scale public database of synthesized, converted, and replayed speech. It was created for the third automatic speaker verification spoofing and countermeasures challenge (ASVSpooF 2019), which was held in 2019. The dataset is designed to help researchers develop and evaluate techniques for detecting spoofing attacks against ASV systems [29]. The ASVspooF 2019 dataset consists of over 13,000 audio recordings, divided into two scenarios such as logical access attacks and physical access attacks. The proposed study considers a logical access attacks dataset as a case study, including genuine and spoofed recordings. The genuine recordings were created by recording the target speakers in a controlled environment with over 20,000 audio recordings, created using various techniques, such as voice conversion and speech synthesis.

2.3. Meta-data construction

This section discusses the process of metadata preparation, which involves extracting essential information from protocol files and organizing it into CSV files for efficient data management and model training. Algorithm 1 is designed that initiates with three inputs: protocol file (P), representing the file containing the required protocol data; audio directory (A), the directory containing audio data; and CSV file (C), signifying the file where the extracted metadata is to be stored. The algorithm's output is the structured list, metadata (M), containing the processed information extracted from the protocol file.

Algorithm 1. Metadata construction for ASV

Input: P (protocol file), A (audio directory), C (CSV file)

Output: M (metadata)

Start

1. Initialize $M = \{ \}$
2. Open protocol file P for reading.
3. For each line p_i in P :
4. Let Parts be the result of splitting the line by whitespace
5. Parts = split(p_i , ' ')
6. //Parts \in [S_Id (speaker identity), F_Id (file identity), ... $Label$]
7. Extract the following information from Parts:
8. For each p_i , extract
 - S_ID \rightarrow Parts[0]
 - F_ID \rightarrow Parts[1]
 - Label \rightarrow Parts[4]
9. Construct AudioPath Ap as:
 - $Ap = A + \text{"flac"} + F_ID$
10. Create a metadata record with:
11. For each p_i , let r_i be a record:
 - $r_i = \{S_Id, F_Id, Label, Ap\}$
12. Append Record to Metadata M
13. $M = M \cup \{r_i\}$
14. Close protocol file
15. Open csv file C for writing
16. Write a header row: S_ID, F_ID, Ap
17. For each r_i in M :
18. Write a row with values
19. writeRow($C, r_i[S_Id], r_i[F_Id], r_i[Label], r_i[Ap]$)
20. Close C
21. Return M

End

An empty list, M , is initially initialized to store the extracted metadata. The algorithm begins by opening the protocol file P for reading. The algorithm performs a series of extraction and processing steps for each line p_i present in the protocol file P . The line is split by whitespace into constituent parts called Parts. Several pieces of information are extracted from these parts, such as speaker identity (S_ID) derived from the first element, file identity (F_ID) from the second element, and label from the fifth element of the parts. Following the extraction, the algorithm constructs the AudioPath. It forms this by concatenating the A with a string "flac/" and the F_ID . Subsequently, a metadata record, denoted as r_i , is created with the extracted S_ID , F_ID , label, and the constructed AudioPath, denoted as Ap . Further, each record is then appended to the metadata list, which accumulates all the metadata records processed from the P . After processing all lines in the P , the file is closed, and the CSV file is opened for writing. A header row is written first into the csv file, containing the columns: 'S_ID', 'F_ID', Ap . Subsequently, for each record r_i present in the metadata M , a new row is written to the CSV file, with values corresponding to the elements of the record (line 19) and upon writing all the metadata records to the CSV file, it is finally closed.

Figures 2 and 3 offer comprehensive visual analyses of bonafide and spoof voice signals, respectively. Figure 2 displays the waveform and spectrogram of a genuine or bonafide voice signal, illuminating its unique temporal and spectral characteristics, which are crucial for ASV systems. In contrast, Figure 3 outlines the characteristics of a spoof or artificially generated voice signal, emphasizing the anomalies and discrepancies in its spectral representation compared to bonafide signals.

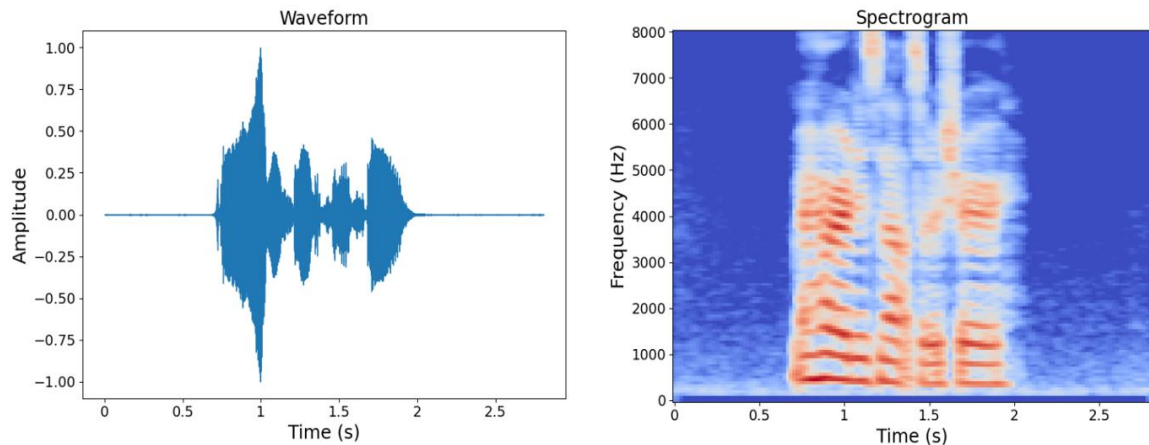


Figure 2. Illustrative depiction of bonafide speaker signal characteristics

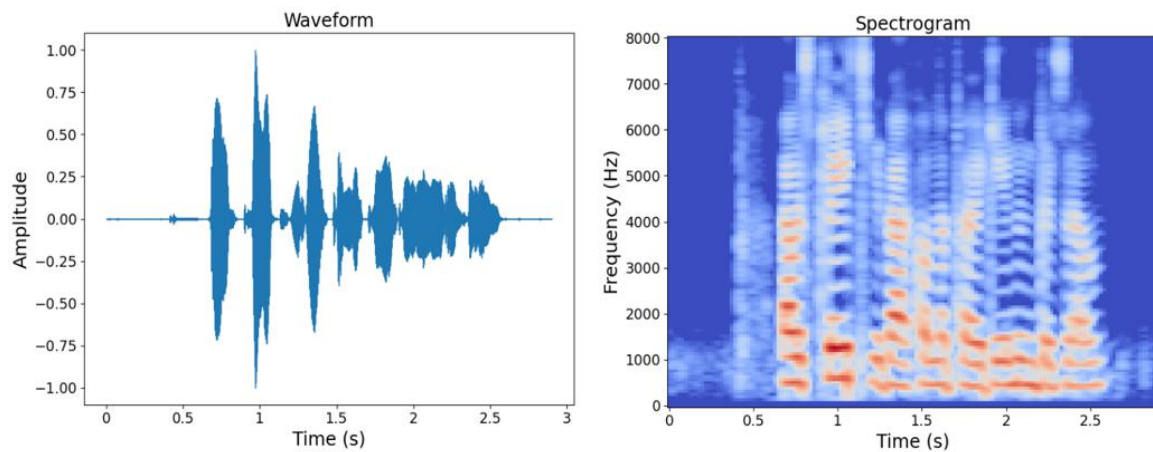


Figure 3. A graphic representation of distinctive spoof voice signal attributes

2. 4. Feature extraction

The feature extraction process is crucial in the architecture of speech-spoofing countermeasures systems. As a transformation step, it converts the original speaker audio data into a structured form, which is advantageous to the detailed analysis and application of machine learning algorithms. Given the inherently discriminative nature of audio signals, capturing the essential features that facilitate accurate discrimination between real (bonafide) and spoofed voices is important. Therefore, the study focuses on extracting three essential features, such as MFCCs, spectral contrast, and tonnetzpitch features that are crucial for distinguishing bonafide and synthetic spoofed speech signals. However, the variability in audio signal duration often led to inconsistent feature lengths. To correct this inconsistency, the introduced method incorporates padding, a technique that aligns feature lengths by extending or truncating them to a predetermined length. A sample visualization and description of the extracted features are depicted in Figures 4 and 5.

Figure 4 displays the two most popular audio features, with subfigure Figure 4(a) representing the short-term power spectrum of an audio signal that captures the spectral characteristics of the signal in a way that approximates the human auditory system's response to sound. This representation showcases the variation of MFCCs over time, providing insights into the audio signal's spectral content and acoustic properties, which are valuable for identifying transient events, speech patterns, and changes in frequency components within the audio signal. Figure 4(b) demonstrates the spectral contrast of the audio signal, a representation that quantifies the difference in amplitude between spectral peaks and valleys in an audio spectrum. This can be particularly useful for distinguishing between different types of audios, as it emphasizes variations in spectral structure, such as the presence of harmonics or formants.

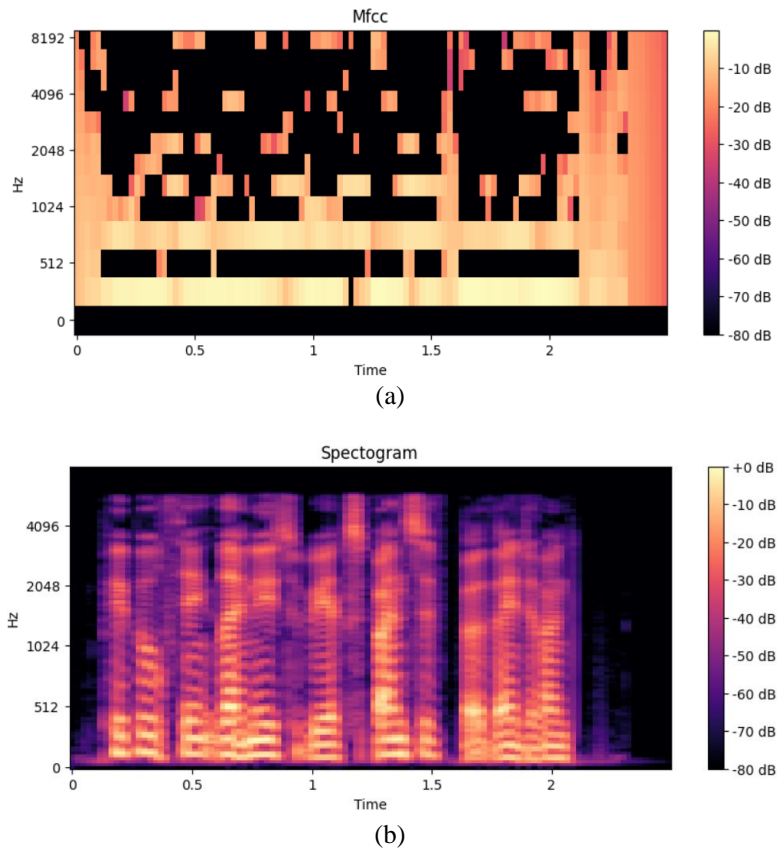


Figure 4. An analytical visualization of (a) MFCC feature attributes and (b) spectrogram representation feature attributes

Figure 5 displays the two critical auditory features; Figure 5(a) displays a spectrogram visualization for audio analysis, a 2D representation of an audio signal where time is plotted on the x-axis, frequency on the y-axis, and the colour intensity represents the magnitude or power of different frequency components at each point in time. This visual representation enables us to observe how the spectral content of the audio evolves over time. In next Figure 5(b), the tonnetz feature of an audio signal is depicted. This feature captures the tonal characteristics of an audio signal, represented as a set of values or coefficients that describe the energy distribution in different tonal regions. Moreover, tonnetz features are sensitive to pitch variations and harmonic content, making them valuable for identifying tonal patterns and differences between genuine and spoofed voices in ASV systems. The extracted features are empirically reviewed, and it is identified that MFCC features, spectral contrast, and Tonnetz features are critical for distinguishing the difference between spoofed and real voices, providing unique insights into the inherent characteristics of a speaker's voice. In contrast, MFCC features, and Mel spectrograms exhibit similar features derived from the input audio data because MFCCs are derived from Mel spectrograms.

Therefore, based on empirical research, this study mainly focuses on MFCC, spectral contrast and Tonnetz characteristics. Principal component analysis (PCA) was deployed on spectral contrast and Tonnetz features to optimize analysis efficiency, aiming to reduce dimensionality while retaining 99% of the explained variation factors. The reason why MFCC is exempted from PCA is based on the empirical observation that its contribution to yield improvement is minimal. The entire computing process involved in extracting and selecting audio features is discussed in Algorithm 2.

Algorithm 2. Feature extraction and selection

Input: A (Set of raw audio data); L_{max} (Maximum allowable feature length)

Outputs: F (Transformed feature matrix); S (Selected feature set)

Start

1. Extract initial feature set $F = \{\text{MFCCs}, \text{Spectrogram}, \text{Spectral Contrast}, \text{Tonnetz}\}$ from A
2. Let $F' = \emptyset$: Initialize the transformed feature matrix
3. Dimensional Consistency:

4. For each feature F_i in F :
5. Compute $\text{num_time_frames} = \text{length}(F_i)$
6. If $\text{num_time_frames} > L_{\max}$:
7. Truncate: $F_i' = F_i[:, : L_{\max}]$
8. Else If $\text{num_time_frames} < L_{\max}$:
9. Extend: $F_i' = \text{Pad}(F_i, (0,0), (0, L_{\max} - \text{num_times_frames})), \text{mode} = \text{'constant'}$
10. Else:
11. $F_i = F_i$
12. Update F' with F_i'
13. Perform analysis on F' to derive discriminative characteristics and similarities.
14. Let $S = \{\text{MFCC}_s, \text{Spectral Contrast}, \text{Tonnetz}\}$ be the set of selected features
15. Apply PCA on feature in S excluding MFCC to retain 99% of the explained variance factor.
16. Update the set S with the features after dimensionality reduction.
17. Return the transformed feature matrix F' and the selected features set S .
18. Construct a feature matrix by concatenating each feature in Set S

End

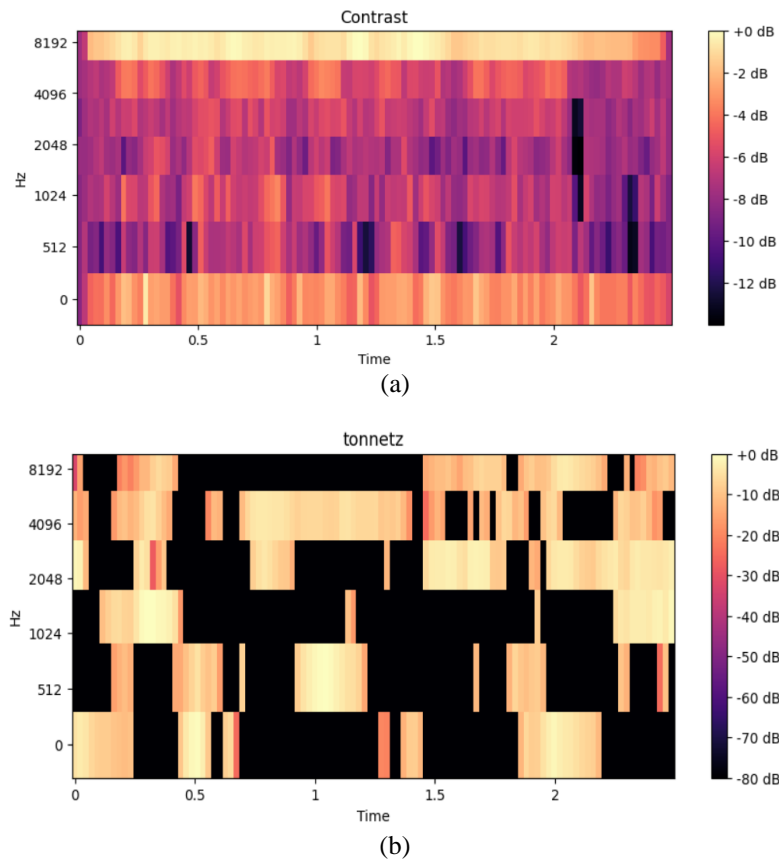


Figure 5. Detailed visual examination of (a) spectral contrast features and (b) tonnetz feature

2.6. Hybrid learning model

The proposed hybrid learning model presents a comprehensive model by leveraging the distinct strengths of both architectures CNN and LSTM networks to adeptly capture and interpret the spectral and temporal intricacies within audio features. The proposed hybrid learning architecture consists of an input layer designed effectively to handle feature vectors. This layer accepts dimensions of shape (2600, 1), where 2600 symbolizes the cumulative length of the MFCCs, spectral contrast, and tonnetz features. The singular channel denotes its unidimensional nature, underscoring the importance of each feature's sequential alignment. Succeeding the input is a convolutional framework consisting of a sequence of one-dimensional convolutional layers.

With their weight-sharing mechanism, these layers attempt to detect and extrapolate spatial relationships within the audio data. In order to ensure computational efficiency without compromising on essential feature information, each convolutional layer is followed by a max-pooling layer, effectively down-sampling the feature dimensions. Building on the spatial patterns extracted by the convolutional layers, the architecture integrates an LSTM layer to understand sequences. The LSTM layer offers the model a sophisticated memory mechanism, empowering it to learn and remember temporal patterns embedded within the audio features. Post the LSTM integration, the modelling phase also adds two fully connected or dense layers. These layers, laden with neurons capable of complex non-linear transformations, further refine the features, preparing them for the final classification task. The study also ensures that the proposed learning model is robust against overfitting issues in the training process by incorporating L2 regularization and dropout techniques.

The final layer of the model is the output layer, which acts as the final decision-making module of the proposed system. It consists of a single neuron with a sigmoid activation function, and it outputs the probability of the input audio being genuine or spoofed. To train the hybrid model effectively, the study used the Adam optimizer, known for its adaptability, with a learning rate initialized at 0.001. The study also used the binary cross-entropy loss function to quantify the model's errors and guide its learning, as it is well-suited for binary classification tasks. The study used two callbacks to ensure consistent and efficient training, i.e., TensorBoard for performance monitoring and ReduceLRonPlateau for dynamic learning rate adjustments. The model was trained over 100 epochs, processing batches of 100 samples each. Figure 6 presents the entire flow chart work for detecting bonafide and voice spoofing attacks using the proposed hybrid spectral-temporal learning model.

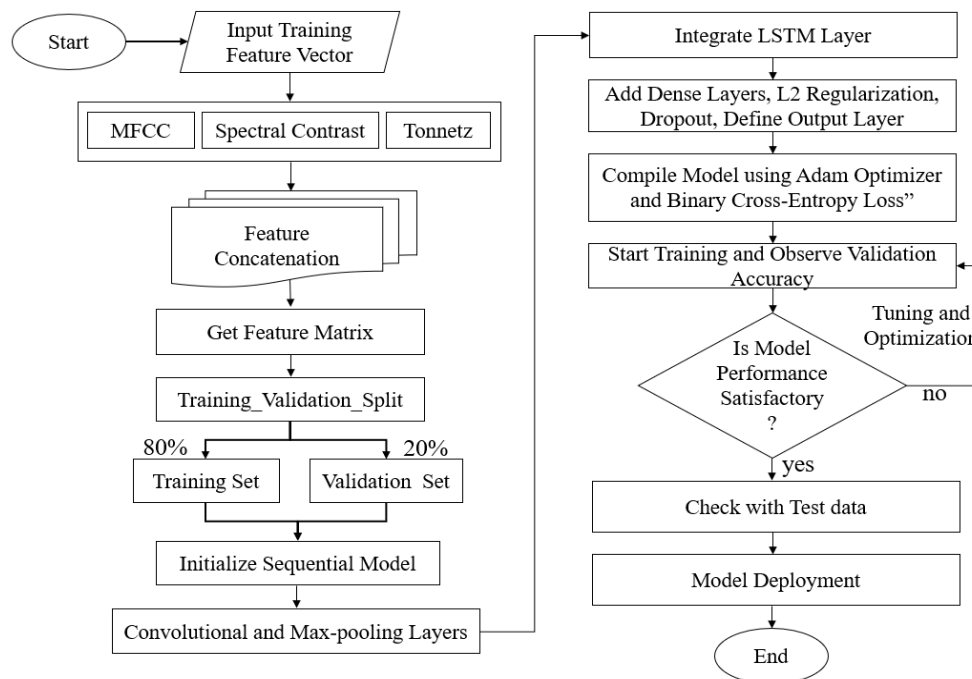


Figure 6. Flowchart of the proposed system for bonafide and spoof signal detection

3. RESULT AND DISCUSSION

The design and development of the proposed system for voice spoofing attack detection is done using Python executed in Anaconda distribution. This section presents the experimental outcome and discusses the performance concerning training accuracy, validation accuracy, confusion matrix, and other classification metrics such as accuracy, precision, recall, F1-score, and EER. Figure 7 displays the model's training performance trends, with Figure 7(a) depicting training and validation accuracy and Figure 7(b) detailing training and validation loss. Analysis of the graph trends reveals a sustained high training accuracy up to 100 epochs, reaching 99.98% and 99.96% for training and validation, respectively. The stable training loss implies minimal errors in distinguishing between genuine and spoofed signals. The frequent invocation of the ReduceLRonPlateau callback, which adjusts the learning rate when validation performance plateaus, indicates that the model was progressively refining towards optimal performance. In Figure 8, the confusion matrix

presents the classification performance of the proposed spectral-temporal learning model for bonafide and spoof classes.

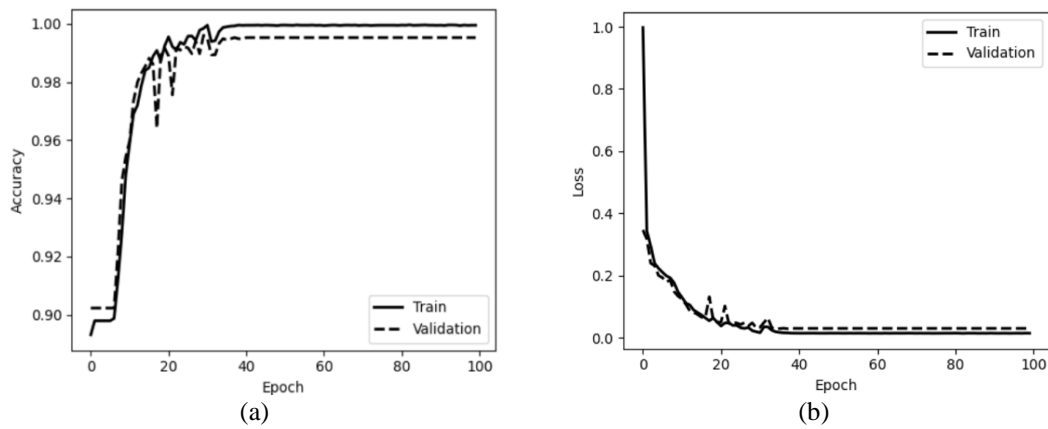


Figure 7. Performance trend of the model's training and validation performance on (a) accuracy and (b) loss

True	bonafide	2443	105
	spoof	35	22261
		bonafide	spoof
		Predicted	

Figure 8. The confusion matrix for the classification model

The interpretation of this confusion matrix is presented in Table 1. A closer analysis of the outcome statistics shows that the model accurately identified 2,443 instances as bonafide out of 2,548 instances in the test dataset, showcasing its ability to discern genuine instances. However, there were 105 instances where bonafide (genuine) was mistaken as spoofed. In the context of spoof instances, the proposed classification model correctly predicted 22,261 samples as spoof signals out of 22,296 test samples, it also falsely categorized 35spoof instances as bonafide.

Table 1. Summary of confusion plot with classification statistics for spoof and bonafide classification

Class label	Total samples	True positive	False negative
Bonafide	2548	2443	105
Spoof	22296	22261	35

The outcome statistics from Table 1 demonstrate the considerable discernment capabilities of the proposed classification model in distinguishing between bonafide and spoofedspeakers. The model demonstrated high precision, accuracy, recall, and F1 scores for speech instances in the test dataset, as shown in Table 2, correctly identifying most instances of bonafide and spoof. In the bonafide class, the model yields impressive results, with a precision of 0.99, recall of 0.96 and F1-score of 0.97. This suggests that the model has a very low false positive rate, and most of the samples classified as bonafide are true positives. With a larger number of samples, the spoof class witnesses a remarkable performance, with no false negatives or positives. When considering the overall outcome, the model's precision is 99.56%, and recall is higher at 99.84%, denoting that the model successfully identifies almost all true instances of each class in the dataset. Subsequently, the F1-score is also high at 99.69%, signifying an exceptional balance between precision and recall. Additionally, EER is 0.014, which is quite low vs the epochs presented in Figure 9.

Table 2. Presents a quantitative evaluation of classification model performance metrics

Total samples	Class	Precision	Recall	F1-score
2548	Bonafide	0.99	0.96	0.97
22296	Spoof	1.00	1.00	1.00
Overall outcome				
Accuracy	Precision	Recall	F1-score	EER
99.44	99.56	99.84	99.69	0.014

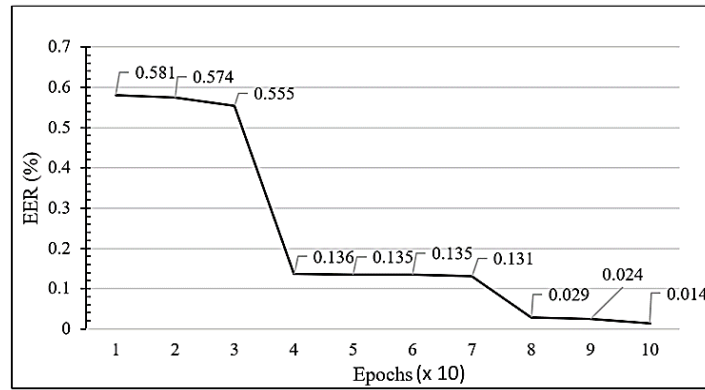


Figure 9. Analysis of EER vs epochs

From Figure 9, it can be observed that the EER starts high at 0.581 in epoch 1, moderately decreases to 0.574 in epoch 2, and further to 0.555 in epoch 3. This indicates that the model is in the early stages of learning, making substantial errors in distinguishing between classes. However, there is an improvement in epochs 4 to 7, with EER values sharply declining from 0.136 to 0.131. It suggests that the model has made significant learnings and optimizations, improving its overall classification accuracy. In the last three epochs, EER reduces dramatically to 0.024 in epoch 8 and remains constant in epoch 9 before achieving the lowest value of 0.014 in epoch 10. The consistent reduction in EER values from epoch 1 to epoch 10 demonstrates the effectiveness of the training process, indicating a consistent learning and adaptation by the model, optimizing its classification capabilities over time. Table 3 provides a comparative analysis in terms of EER against two similar existing methodologies.

Table 3. Presents comparative analysis in terms of EER

Existing 1 [15]	Existing 2 [27]	Proposed
2.19	0.23	0.014

The comparative analysis presented in Table 3, in terms of EER, clearly demonstrates the superiority of the proposed model over the existing models [15], [27]. Mittal and Dua [15] demonstrated an EER of 2.19, which is considerably higher than the existing model [27] and proposed system. In Mittal and Dua [15] one-class learning method is implemented to identify genuine speech representations and introduce an angular margin without data augmentation. However, this approach has limitations in terms of scalability and adaptability to diverse spoofing scenarios. In contrast, Adiban *et al.* [27] adopted autoencoder and Siamese networks, which resulted EER of 0.23%, demonstrating enhanced performance compared to [15] in differentiating legitimate from spoofed voices. In the case of the proposed system has achieved a quite low EER of 0.014, indicating a superior level of accuracy and reliability in preventing voice spoofing attacks, suggesting the model has achieved a highly optimized and accurate state, making it reliable for classifying bonafide and spoof classes. The proposed system achieves good performance due to the usage of hybrid spectro-temporal features combined with a sophisticated learning model.

4. CONCLUSION

This study proposes a novel learning method model to solve the problem of logical speech spoofing attacks on ASV systems. The model combines the learning capabilities of CNN and LSTM networks, which are good at learning from spectral and temporal data. The proposed method shows remarkable adaptability in

detecting changes in the speaker's voice and distinguishing authentic from spoofed voices. The uniqueness of the proposed learning model is that it achieves learnability, efficiency, and practicality. It extracts multifaceted features from speech signals, thereby covering different speech attributes to facilitate a comprehensive analysis of speech patterns. The entire analysis of the experimental results highlights the significant advantages of this approach over existing spoofing attack countermeasure solutions. Performance metrics validate research findings emphasizing system robustness and reliability. The proposed system achieves an accuracy of 99.44%, precision of 99.56%, recall of 99.84%, F1 score of 99.69%, and EER of 0.014, outperforming existing similar work. The proposed model will be extended to detect replay attacks with only minor changes in feature extraction and training parameter tuning.




REFERENCES

- [1] A. Sarkar and B. K. Singh, "A review on performance, security and various biometric template protection schemes for biometric authentication systems," *Multimedia Tools and Applications*, vol. 79, no. 37–38, pp. 27721–27776, Oct. 2020, doi: 10.1007/s11042-020-09197-7.
- [2] W. Yang, S. Wang, N. M. Sahri, N. M. Karie, M. Ahmed, and C. Valli, "Biometrics for internet-of-things security: a review," *Sensors*, vol. 21, no. 18, Sep. 2021, doi: 10.3390/s21186163.
- [3] O. S. Asaolu, C. Folorunso, and O. Popoola, "A review of voice-base person identification: state-of-the-art," *Journal of Engineering Technology*, vol. 3, no. 1, pp. 36–57, 2019, doi: 10.20370/2cdk-7y54.
- [4] C. Simon and M. Rajeswari, "Voice-based virtual assistant with security," in *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, Mar. 2023, pp. 822–827, doi: 10.1109/ICEARS56392.2023.10085043.
- [5] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, Aug. 2021, doi: 10.1016/j.neunet.2021.03.004.
- [6] A. Mittal and M. Dua, "Automatic speaker verification systems and spoof detection techniques: review and analysis," *International Journal of Speech Technology*, vol. 25, no. 1, pp. 105–134, Mar. 2022, doi: 10.1007/s10772-021-09876-2.
- [7] P. N. Palsapure, R. Rajeswari, and S. K. Kempegowda, "Enhancing speaker verification accuracy with deep ensemble learning and inclusion of multifaceted demographic factors," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 6, pp. 6972–6983, Dec. 2023, doi: 10.11591/ijece.v13i6.pp6972-6983.
- [8] C. B. Tan *et al.*, "A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction," *Multimedia Tools and Applications*, vol. 80, no. 21–23, pp. 32725–32762, Sep. 2021, doi: 10.1007/s11042-021-11235-x.
- [9] H. A. Patil and M. R. Kamble, "A survey on replay attack detection for automatic speaker verification (ASV) system," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Nov. 2018, pp. 1047–1053, doi: 10.23919/APSIPA.2018.8659666.
- [10] M. FarrÚs, "Voice disguise in automatic speaker recognition," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–22, Jul. 2019, doi: 10.1145/3195832.
- [11] C. Yan, X. Ji, K. Wang, Q. Jiang, Z. Jin, and W. Xu, "A survey on voice assistant security: attacks and countermeasures," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–36, Apr. 2023, doi: 10.1145/3527153.
- [12] P. Partila, J. Tovarek, G. H. Ilk, J. Rozhon, and M. Voznak, "Deep learning serves voice cloning: how vulnerable are automatic speaker verification systems to spoofing trials?," *IEEE Communications Magazine*, vol. 58, no. 2, pp. 100–105, Feb. 2020, doi: 10.1109/MCOM.001.1900396.
- [13] M. Mcuba, A. Singh, R. A. Ikuesan, and H. Venter, "The effect of deep learning methods on deepfake audio detection for digital investigation," *Procedia Computer Science*, vol. 219, pp. 211–219, 2023, doi: 10.1016/j.procs.2023.01.283.
- [14] R. Magazine, A. Agarwal, A. Hedge, and S. R. M. Prasanna, "Fake speech detection using modulation spectrogram," in *Speech and Computer*, Springer International Publishing, 2022, pp. 451–463.
- [15] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021, doi: 10.1109/LSP.2021.3076358.
- [16] M. Dua, C. Jain, and S. Kumar, "LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 4, pp. 1985–2000, Apr. 2022, doi: 10.1007/s12652-021-02960-0.
- [17] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, "The PartialSpoof database and countermeasures for the detection of short fake speech segments embedded in an utterance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 813–825, 2023, doi: 10.1109/TASLP.2022.3233236.
- [18] J. Xue *et al.*, "Learning from yourself: a self-distillation method for fake speech detection," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10096837.
- [19] Z. Lei, Y. Yang, C. Liu, and J. Ye, "Siamese convolutional neural network using gaussian probability feature for spoofing speech detection," in *Interspeech 2020*, Oct. 2020, pp. 1116–1120, doi: 10.21437/Interspeech.2020-2723.
- [20] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," in *Interspeech 2020*, Oct. 2020, pp. 1101–1105, doi: 10.21437/Interspeech.2020-1810.
- [21] A. Javed, K. M. Malik, H. Malik, and A. Irtaza, "Voice spoofing detector: A unified anti-spoofing framework," *Expert Systems with Applications*, vol. 198, Jul. 2022, doi: 10.1016/j.eswa.2022.116770.
- [22] S. K. Kempegowda, R. Rajeswari, L. Satyanarayana, and S. Matada Basavarajaiah, "Hybrid features and ensembles of convolution neural networks for weed detection," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 6, pp. 6756–6767, Dec. 2022, doi: 10.11591/ijece.v12i6.pp6756-6767.
- [23] I.-Y. Kwak *et al.*, "Voice spoofing detection through residual network, max feature map, and depthwise separable convolution," *IEEE Access*, vol. 11, pp. 49140–49152, 2023, doi: 10.1109/ACCESS.2023.3275790.
- [24] I.-Y. Kwak, S. Choi, J. Yang, Y. Lee, S. Han, and S. Oh, "Low-quality fake audio detection through frequency feature masking," in *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, Oct. 2022, pp. 9–17, doi: 10.1145/3552466.3556533.
- [25] J. Guo, Y. Zhao, and H. Wang, "Generalized spoof detection and incremental algorithm recognition for voice spoofing," *Applied*




- Sciences*, vol. 13, no. 13, Jun. 2023, doi: 10.3390/app13137773.
- [26] K. M. Malik, A. Javed, H. Malik, and A. Irtaza, "A light-weight replay detection framework for voice controlled IoT devices," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 982–996, Aug. 2020, doi: 10.1109/JSTSP.2020.2999828.
- [27] M. Adiban, H. Sameti, and S. Shehnepoor, "Replay spoofing countermeasure using autoencoder and siamese networks on ASVspoof 2019 challenge," *Computer Speech & Language*, vol. 64, Nov. 2020, doi: 10.1016/j.csl.2020.101105.
- [28] X. Wang *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, Nov. 2020, doi: 10.1016/j.csl.2020.101114.
- [29] A. Nautsch *et al.*, "ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, Apr. 2021, doi: 10.1109/TBIOM.2021.3059479.

BIOGRAPHIES OF AUTHORS






Pranita Niraj Palsapure    is working presently as an Assistant professor in the Department of Electronics and Communication Engineering at Acharya Institute of Technology, Bangalore, Karnataka. She is pursuing her Ph. D. under Visvesvaraya Technological University, Belgavi, Karnataka, India and M. Tech from Nagpur University, Maharashtra in 2007. Her area of research is Speech processing and Machine learning. She is a member of ISTE. She can be contacted at email: pranitanirajpalsapure@gmail.com.






Dr. Rajeswari    is associated with Acharya Institute of Technology, Bangalore, India as Professor in the Department of Electronics and Communication Engineering. She has completed her Ph. D. in the field of speech processing. Her areas of interest include speech processing, AI, computer vision and application in the field of healthcare and agritech. She is CMI level 5 certified in Management and Leadership under UKIERI. She can be contacted at email: rajeswari@acharya.ac.in.



Sandeep Kumar Kempegowda    is presently working as Assistant professor in Department of Electronics and Communication Engineering at Acharya Institute of Technology, Bangalore, Karnataka. He is pursuing his Ph. D. under Visvesvaraya Technological University, Belgavi, Karnataka, India, M. E. (ECE) from Bangalore University, Karnataka in 2010. His area of research is image processing, computer vision, machine learning and embedded systems. He is a member of ISTE. He can be contacted at email: sandy85gowda@gmail.com.



Kumbhar Trupti Ravikumar    is presently working as Assistant Professor in the Department of Electronics and Communication Engineering at Acharya Institute of Technology, Bangalore, Karnataka. She pursued M. E. (E & Tc). Her area of interest is digital signal processing, embedded systems. She is a member of IETE. She can be contacted at email: trups304@gmail.com.