Deep feature synthesis approach using selective graph attention for replay attack voice spoofing detection

Pranita Niraj Palsapure, Rajeswari, Sandeep Kumar Kempegowda
Department of Electronics and Communication Engineering, Acharya Institute of Technology, Visvesvaraya Technological University, Belagavi, India

Article Info

Article history:

Received Jan 17, 2024
Revised Mar 26, 2024
Accepted Apr 17, 2024

Keywords:

Automatic speaker verification
voice spoofing
Deep learning
Graph attention
Replay attack detection
Spectral-temporal feature

ABSTRACT

As voice-based authentication becomes increasingly integrated into security frameworks, establishing effective defenses against voice spoofing, particularly replay attacks, is more crucial than ever. This paper presents a novel comprehensive framework for replay attack detection that leverages the integration of advanced spectral-temporal feature extraction and graph-based feature processing mechanisms. The proposed system presents the design of a waveform encoder and a novel temporal residual unit for spectral and temporal feature extraction in synchronous. Further, an approach of selective attention graph followed by multi-scale feature synthesis is employed to retain precise and spoof indicative feature vectors at the classification layer. The proposed method addresses the significant challenge of distinguishing genuine speech from replayed recordings. The validation of the proposed model is done on the ASVSpoof2019 dataset to demonstrate the efficacy of the proposed approach. The proposed system outperforms existing methods, achieving a lower equal error rate (EER) of 0.015 and a reduced tandem detection cost function (t-DCF) of 0.503. The comparative outcome exhibits the robustness of the method in identifying replay attacks.

Corresponding Author:

Pranita Niraj Palsapure
Department of Electronics and Communication Engineering, Acharya Institute of Technology
Visvesvaraya Technological University
Bangalore, India
Email: pranitanirajpalsapure@gmail.com

1. INTRODUCTION

Voice-based authentication systems use an individual's unique voice characteristics and have become an integral part of modern security systems ranging from personal device security to enterprise data security [1], [2]. However, due to its higher adoption also brought new security vulnerabilities, especially voice spoofing attacks. Automatic speaker verification (ASV) systems use voice biometrics to verify identity by analyzing speech characteristics such as pitch and tone [3]. However, they are increasingly exposed to the risk of voice spoofing, where attackers copy or manipulate voice signals to breach security [4]. Among speech spoofing methods, replay attacks pose a particularly difficult challenge [5], [6]. In replay attacks, malicious actors use recordings of legitimate users, a strategy that is both relatively easy and efficient, thus becoming a preferred method for compromising the reliability of ASV systems [3], [7]. The primary challenge in countering replay attack is the minute difference between real speech and spoofed speech, which often undetectable by traditional ASV systems [8]. This difficulty is further intensified by rapidly evolving recording and playback technologies that create high-quality analogue audio that is indistinguishable from real speech [9], [10]. Hence, detecting replay voice attacks faces several potential challenges, including high accuracy in feature discrimination and the need for systems to adapt to evolving techniques [11], [12].

In recent state-of-art works, different researchers have carried out many works, where deep learning approaches and their hybridization is done to build an effective detection model but at the cost of computationally intensive modelling. The researchers in the study of Gong *et al.* [13] raised critical concerns about the security of ASV systems against evolving replay attacks. The authors have presented a new replay attack dataset named realistic replay attack corpus for voice-controlled systems (ReMASC), developed specifically for assessing vulnerabilities in ASV systems against modern replay attacks on text-dependent systems under varying recording and playback conditions. Wu *et al.* [14] identified the significant vulnerability of text-dependent speaker verification systems to replay attacks. Using a similarity score, they presented an anti-spoofing technique that compares presented speech samples to previously stored ones. Li *et al.* [15] focus on overcoming the over-fitting problem in replay detection, often caused by variability factors in speech signals. A frequency warping approach is then proposed and successfully tested on the ASV-spoof 2017 database, demonstrating its effectiveness in reducing over-fitting and enhancing replay attack detection. Alegre *et al.* [16] critically reevaluate the risks of spoofing attacks on ASV systems and highlight the greater risk of replay attacks due to their simplicity and lack of required technical expertise. Through comprehensive testing against six different ASV systems, including an advanced iVector-probabilistic linear discriminant analysis (PLDA) system, the study demonstrates that low-effort replay attacks result in higher false acceptance rates compared to more complex spoofing methods. In the study of Xue *et al.* [17], an iterative knowledge distillation method is adopted for fake speech detection where a deep network as the instructor model to guide multiple shallow classifiers by minimizing feature differences. Lei *et al.* [18] developed a method to detect known and unknown spoofing attacks using 1-D convolutional neural network (CNNs) and, Siamese CNN and Gaussian mixture model (GMM) components to capture both local and global speech features. Wu *et al.* [19] 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.* [20] developed a framework that uses co-occurrence patterns and cepstral coefficients to effectively detect distortions and artefacts induced by different spoofing methods, providing comprehensive protection against even complex spoofing attacks. Kwak *et al.* [21], [22] developed new models that are more efficient and robust to unseen spoof attacks. Guo *et al.* [23] 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. Saranya *et al.* [24] presented a method for detecting replay attacks by analyzing reverberation and channel information from non-voiced segments of speech identified using a voice activity detector. The approach utilizes multiple feature representations to capture residual vocal tract information in these segments, employing Gaussian mixture models to create baseline systems for evaluation. Kemanth *et al.* [25] proposed solution that adopts CNNs, leveraging their powerful feature extraction capabilities to identify characteristics indicative of replay attacks. Through this approach, the researchers demonstrate a significant improvement in the system's ability to discern genuine speech from replayed recordings, showcasing the effectiveness of CNNs in enhancing the security of speaker verification systems against such sophisticated threats.

Hence, there are many works in the literature for designing an efficient detection system for replay attacks, which is fraught with challenges. Despite numerous studies in the context of replay attack detection, a significant research gap remains. Most existing works primarily focus on extracting features like mel-frequency cepstral coefficients (MFCC) or mel-spectrogram to train deep learning models. However, this approach overlooks other potentially valuable features, such as tonnetz and spectral contrast. Though less explored in the literature, these features could significantly enhance the detection process. Although CNNs are prevalently used for their ability to capture spectral features, their capacity to process temporal dependencies remains inadequate. However, the recent trends show the usage of integrated and hybrid approaches but at the cost of higher computational cost and lack of optimization in the system design. Replay attack detection fundamentally involves discerning the differences in frequency attributes between genuine and replayed speeches. However, distinguishing these features is a complex task, often hindered by the evolving nature of attack strategies and the limitations of current ASV systems [26]. This can be better understood by the different replay adversarial scenarios shown in Figure 1 against the ASV system. In scenario Figure 1(a) the adversary captures genuine speech directly from the target speaker using a recording device. This speech is then replayed into the ASV system, with potential alterations to incorporate environmental acoustics in an attempt to deceive as the legitimate user. Figure 1(b) depicts an alternative attack vector where the attacker utilizes a previously recorded or acquired digital audio file, bypassing the need for real-time capture. The file is played back to the ASV system, challenging the system's ability to discern its authenticity. While in the third scenario shown in Figure 1(c) represents an advanced technique where the attacker employs a vocal tract emulator. This approach aims to simulate the nuances of human speech production and the corresponding environmental acoustics, thus enhancing the replayed audio's authenticity to deceive the ASV system.

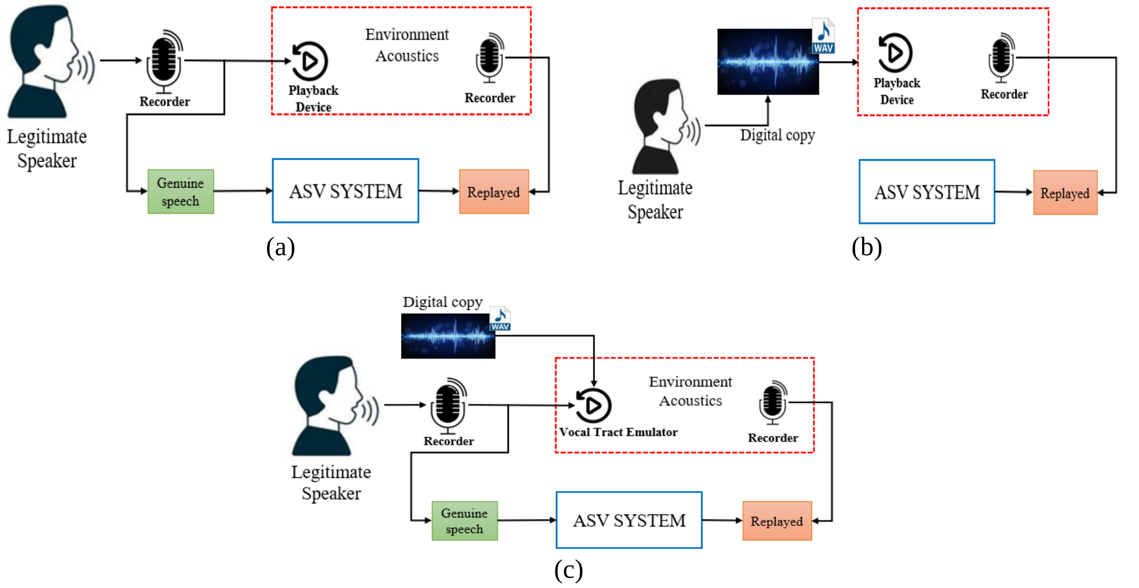

Figure 1. Different replay attack scenarios targeting ASV systems: (a) direct recording replay attack, (b) digital copy replay attack, and (c) digital copy replay attack with vocal tract emulation

These scenarios demonstrate the evolving threat scenario to ASV systems and the demand for designing robust detection mechanisms that can robustly identify the relevant artefacts (device artefacts, and speaker identity), while, at the same time, ignoring variability introduced by the other factors (environmental noise) to generalize well to unknown scenarios. This necessitates using a feature representation with high spectral information to capture details present in spectral regions as well as temporal dependencies that contain discriminative information as an indication of replay attacks. Moreover, the model should also be able to selectively attend to these regions so that it does not overfit the other inessential variability factors.

Therefore, this paper proposes a comprehensive framework specifically designed to counteract the complexities inherent in replay attacks. This framework precisely analyzes acoustic features extracted from raw audio waveforms, employing advanced learning schemes and graph-based analysis techniques to harness the complementary strengths of spectral analysis temporal dynamics. The prime aim of the proposed study is to present a dynamic algorithm capable of adjusting to novel spoofing techniques, thereby ensuring robust defense mechanisms. By examining both spectral and temporal aspects, the proposed approach aims to discern the subtle distinctions between genuine and replayed audio. The design of the proposed system is carried out in such a way that it attempts to achieve a balance between high detection accuracy with manageable computational demands to be feasible for practical deployment. The key contribution of this paper is highlighted as follows:

- This paper introduces an advanced feature representation scheme from raw audio signals augmenting models with spectral and temporal information to capture discriminative replay indications effectively.
- The proposed framework leverages the waveform encoder module that focuses on learning band-pass filters to capture critical frequency components for fine-grained spectral analysis.
- The study has also proposed the implementation of temporal residual units (TRU) to process the temporal aspects of the audio signal.
- A selective attention graph (SAG) layer dynamically weights spectral-temporal regions, preventing overfitting by selectively concentrating on relevant artifacts while disregarding non-essential variability, ensuring robust generalization across diverse scenarios.

The remainder of the manuscript is structured as follows: section 2 outlines the system design methodology, detailing the role of each component and the implementation procedure for detecting replay attacks. Section 3 presents the results and discussion, providing insights into the effectiveness of the proposed framework. Finally, section 4 concludes the paper by summarizing its contributions and discussing implications for future research.

2. METHOD

The proposed study presents a novel computational framework, a multi-layered system designed to analyze audio waveforms, extract spectral-temporal features, process key replay indicative features, and decide whether the audio has been tampered with or is authentic. The proposed framework is designed to process directly raw audio waveforms from the input audio signals, as this choice facilitates direct engagement with the intrinsic properties of the speech signal, bypassing the need for feature extraction typical in conventional signal processing. Unlike many existing works, the proposed study adopts an approach of graph theory in the deep learning model where audio data is processed in a graph, where nodes represent segments and features of the audio, and edges denote relationships or dependencies between these segments. The proposed system also leverages an application of an attention mechanism, making the proposed system focus on analyzing both spectral (frequency-related) and temporal (time-related) aspects of audio data. The attention mechanism enables the model to focus on the most relevant parts of this data, which are crucial for detecting spoofing attacks. The study also introduces a significant feature extraction operation within the graph structure, ensuring that only the most relevant or distinctive features are considered when the system decides whether an utterance is spoofed or bonafide. Figure 2 presents the schematic architecture of the proposed system following various specialized and technologically advanced computing modules strategically integrated in a highly synchronized manner.

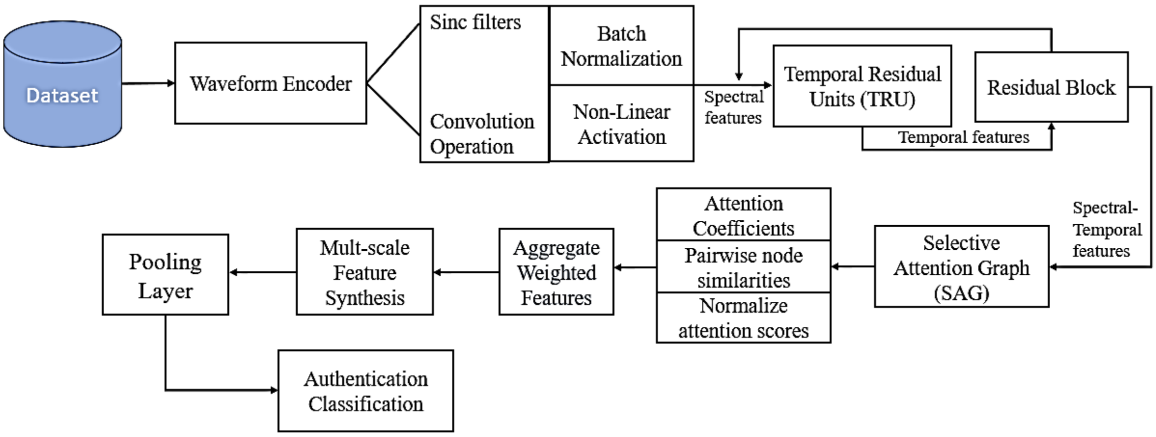
Dataset
Waveform Encoder
Sinc filters
Convolution Operation
Batch Normalization
Non-Linear Activation
Spectral features
Temporal Residual Units (TRU)
Residual Block
Temporal features
Spectral-Temporal features
Selective Attention Graph (SAG)
Attention Coefficients
Pairwise node similarities
Normalize attention scores
Aggregate Weighted Features
Mult-scale Feature Synthesis
Pooling Layer
Authentication Classification

Figure 2. Illustrating high-level architecture of the proposed system for replay attack detection

The architecture of the proposed system shown in Figure 2 adopted a supervised learning approach in the training. The system first implements a waveform encode module, which takes input audio signal from the dataset and processes it to extract high-level spectral features. This module employs a data-driven strategy with 2D-CNN and parametrized Sinc functions that act as band-pass filters to capture robust frequency information, offering a high-level spectral attribute without reliance on handcrafted feature extraction methods. The waveform encoder also uses batch normalization after convolution operation to ensure the input data distribution for each mini-batch during training stays consistent. Following batch normalization, a non-linear activation function introduces non-linearity, enabling the network to learn complex patterns and maintaining a self-normalizing property that promotes a stable learning process. The second module, TRU, is designed to process and enhance spectral features by integrating temporal dynamics, enabling the model to comprehend complex speech patterns that evolve over time. The TRU module integrates long short-term memory (LSTM) with skip-connection blocks to capture time-dependent patterns and prevent the vanishing gradient problem, preserving rich spectral and temporal information. The next important module of the proposed system is the SAG module, which represents the graph attention mechanism. The SAG module computes attention coefficients that weigh the features' importance, allowing the model to focus on the most informative aspects of the input data. Afterwards, the proposed study implements a multi-scale feature synthesis function, allowing the model to synthesize features at different scales, thus capturing both local details and global contextual information, highlighting traits indicative of replay attacks. Before classification, a pooling layer is utilized to reduce the feature maps' dimensionality and make the network more computationally efficient. The final layer authentication classification extracts a final decision regarding the authenticity of the audio signal. This module employs a linear layer that transforms the pooled features into a decision space, followed by a Soft-max function that yields a probability distribution over potential class i.e. bonafide or spoofed audio.

2.1. Dataset description

This research study utilizes the ASVSpoof2019 dataset [27], a robust and extensive collection of vocal samples designed for spoofing and countermeasure analysis. Originating from the voice cloning toolkit (VCTK) corpus, this dataset encompasses a variety of artificial, altered, and replayed voice samples from 107 speakers, comprising 46 males and 61 females. ASVSpoof 2019 is structured into three primary segments: training, development, and evaluation. The training and development sets are subjected to 20 speakers, divided between targets and non-targets. The evaluation set, however, includes 67 speakers, 48 of them targets and 19 non-targets. This dataset focuses on assessing the impact of spoofing countermeasures on ASV systems, highlighted by adopting the tandem detection cost function (t-DCF) as a critical metric. This dataset introduces two principal spoofing scenarios:

- Logical access: this scenario simulates attacks directly targeting the ASV system, typically involving synthetic speech or voice conversion. These attacks occur without the influence of acoustic propagation or specific microphone characteristics.
- Physical access: both legitimate and spoofed speeches are considered to traverse a physical space before being captured by the system's microphone. A critical aspect of PA is replaying attacks, where previously recorded legitimate attempts are replayed in the same environment.

Table 1 demonstrates the distribution of speakers and samples across the training, development, and evaluation subsets. The training subset comprises 8 male and 12 female speakers, with 48,600 spoofed and 5,400 bonafide samples. The development subset reflects the gender distribution of the training subset, containing a similar number of spoofed and bonafide samples. The evaluation subset is more extensive as it consists large sample of 30 male and 37 female speakers and a larger pool of 13,4630 spoofed and 18,089 bonafide speech samples. This diverse and comprehensive dataset is instrumental in evaluating the effectiveness of the proposed spoofing detection system under various scenarios.

Table 1. The demonstration of dataset statistics

Subset	Speakers		Class	
	Male	Female	Spoof	Bonafide
Train	8	12	48,600	5400
Dev	8	12	24,300	5400
Eval	30	37	13,4630	18,089

2.2. Waveform encoder

The waveform encoder is the first critical component in the proposed system, responsible for processing raw audio signals. Its primary function is to extract meaningful spectral features from these raw waveforms, essential for the subsequent stages of spoofing detection. The proposed waveform encoder consists of SincNet [28], a specialized type of CNN that utilizes parametrized Sinc functions to perform convolution operations directly on the raw audio waveform. This approach offers a more efficient and interpretable method for feature extraction compared to traditional CNN. The convolution operation in SincNet can be mathematically represented as follows,

The raw input audio signal can be considered as $x[t]$, where t represents discrete time steps. This signal is a time-domain representation of the audio waveform. The SincNet layer within the waveform encoder is designed to process the input $x[t]$ through a series of parametric band-pass filters defined using the Sinc function, acting as an idealized band-pass filter. For a given filter$i$the impulse response is numerically expressed as follows:

$$g[n, f_{1i}, f_{2i}] = 2f_{2i} \times sinc(2\pi f_{2i} n) - 2f_{1i} \times sinc(2\pi f_{1i} n) \quad (1)$$

$$y_i[t] = x[t] * g_i[t, f_{1i}, f_{2i}] \quad (2)$$

In (1) f_1 and f_2 are the lower and upper cut-off frequencies of the filter. The convolution of the input signal with the i-*th* filter is expressed in (2) where, $*$denotes the convolution operation and $y_i[t]$is the output of *i-th* filter. After the convolution operation, batch normalization is applied, followed by non-linear activation. The operation of batch normalization over the obtained feature map can be numerically expressed as (3):

$$\widehat{y_i}[t] = \frac{y_i[t] - \mu_{batch}}{\sqrt{\sigma_{batch}^2 + \epsilon}} \quad (3)$$

Where μ_{batch} and σ^2_{batch} are the mean and variance of the batch, respectively, and ϵ is a small constant to prevent division by zero. This step normalizes the output of each filter across the batch, enhancing the stability and efficiency of the network. To enable the network to capture complex patterns in the data, a non-linear activation function is applied to the batch-normalized output as expressed as (4):

$$z_i[t] = \lambda \begin{cases} \hat{y_i}[t] \; if \; \hat{y_i}[t] > 0 \\ \alpha \times \left(e^{\hat{y_i}[t]} - 1\right) \; otherwise \end{cases} \tag{4}$$

Where, λ and α are predefined parameters of the non-linear activation function. Upon completion of the convolution, batch normalization, and non-linear activation processes, these spectral features are represented as a multi-dimensional tensor, referred to as the spectral feature map tensor. If the SincNet layer employs N filters and the input audio signal $x[t]$, has a length of T discrete time steps, the spectral feature map tensor resulting from the SincNet layer is a NxT matrix. Each element in this tensor denoted as $Z_{i,t}$ corresponds to the activated output of the i-*th* filter at time step t. Thus, for the entire set of N filters, the spectral feature map tensor is represented as (5):

$$Z = \begin{bmatrix} Z_{1,1} & Z_{1,2} & \cdots & Z_{1,T} \\ Z_{2,1} & Z_{2,2} & \cdots & Z_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{N,1} & Z_{N,2} & \cdots & Z_{N,T} \end{bmatrix} \tag{5}$$

Each row in the tensor Z corresponds to the output from one of the$N$SincNet filters across all time steps, encapsulating a specific frequency band's information extracted from the audio signal. Each column in the tensor represents the combined filter outputs at a particular time step, providing a comprehensive spectral representation at that instant. The effectiveness of this module lies in its ability to learn from the raw waveform directly, thus preserving the natural characteristics of the audio while extracting crucial spectral features. These features then serve as the foundation for the subsequent modules in the system, where they are further analyzed for spoofing detection.

2.3. Temporal residual unit module

The TRU is an essential component of the proposed system, designed to capture the temporal dynamics and dependencies inherent in audio signals. It is adept at processing the spectral features extracted by the waveform encoder, further refining these features to emphasize time-based patterns crucial for distinguishing genuine from spoofed audio. The TRU comprises LSTM layers followed by modified residual blocks or skipped connections. The LSTM layers capture temporal dependencies, while the residual blocks ensure the preservation and enhancement of both spectral and temporal information. The output of the TRU is an enriched feature representation that encapsulates both spectral and temporal characteristics of the input audio. The design consideration of TRU includes the following considerations:

- The output of SincNet, a set of feature maps, is reshaped to form a sequence suitable for LSTM processing. This reshaping is crucial to align the spectral features for temporal analysis.
- The LSTM layer takes the reshaped SincNet output and extracts temporal features, capturing time-dependent patterns in the audio signal.
- To enable the residual connection, the output of the LSTM (both feature dimensions and sequence length) must match its input. This ensures seamless addition of the LSTM output back to its input.
- A linear transformation is applied for alignment if there is a mismatch between the LSTM output and the input dimensions. This step is essential for maintaining the integrity of the residual connection.
- The residual block processes the LSTM output to enhance temporal features while preserving spectral information. The block integrates the original LSTM output with the processed output to enrich the feature set.
- It is important to note that the input to the LSTM, which forms the basis for the residual connection, is not the raw output of SincNet but rather its reshaped or processed form suitable for LSTM processing.

Figure 3 shows the implementation of the TRU module, which takes input as a spectral feature map tensor Z from the Waveform Encoder. The LSTM processes the spectral feature map Z in a sequence-to-sequence manner. The output of the LSTM layer can be represented as H=[$h_1, h_2, h_2 \cdots h_T$], where each $h_t \in \mathbb{R}^M$ is the hidden state of the LSTM at time step t. Each residual block within the TRU applies a series of transformations to the LSTM output H. The study denotes this transformation function within the residual block as F(H, Q_{res}), where, Q_{res} represents the parameters of the block.

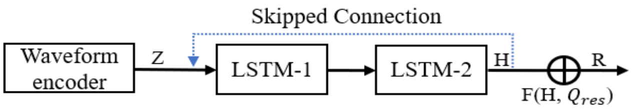

Figure 3. Illustrating schema of TRU module

The output of the residual block is given by R=H+F(H, Q_{res}), where, R is the resultant feature representation with both spectral and temporal features enhanced. This enhanced fearture plays an important role in improving generalization capability of model throughout entire training process and also it allows for the effective learning and retention of significant temporal patterns over extended sequences. By integrating LSTM layers alongside residual blocks, the TRU integrates these patterns with spectral features, thereby improving the overall capability of the system to distinguish between genuine and spoofed audio.

2.4. Selective attention graph module

The SAG module employs an attention mechanism within a graph-based framework, enabling the model to focus selectively on the most pertinent features for spoofing detection. The subtleties between genuine and replayed speech are often minute. The traditional methods often overlook these minute distinctions, leading to less effective spoofing detection. The proposed SAG addresses this limitation by employing an attention layer, which facilitates focused processing of the most relevant features. The attention mechanism in this layer focuses on identifying the most relevant features within the input it receives (containing both spectral and temporal information). It assigns higher weights to more important features for distinguishing genuine speech from replay attacks. This process is based on the relationships and structure of the features as they are represented in a graph format. Each node in the SAG represents a spectral-temporal feature, and the edges signify the relationships or dependencies between these features. By computing attention coefficients, the SAG dynamically adjusts the influence of each node, allowing the network to prioritize features most indicative of spoofing activities. This process enhances the model's sensitivity to relevant cues and suppresses irrelevant noise or distortions that may lead to false detections. The output of the SAG is an aggregated feature set for each node, where the features of neighboring nodes are combined in a weighted manner based on the computed attention scores. This aggregation allows the model to emphasize more relevant features for the detection task. The algorithm-1 describes the implementation and core operation of the SAG module.

Algorithm 1. Feature refinement with selective graph attention

```
Input: Feature Tensor $R \epsilon \mathbb{R}^{B \times L \times D}$, where B is batch size, L is sequence length and D is feature-dimension
Output: Refined feature tensor $Y$
Start
1.   Define linear transform layers for attention computation, $W_{att} \epsilon \mathbb{R}^{D \times D^1}$ and bias $b_{att}$.
2.   Define projection layers $W_{proj}^{att}$,$W_{proj}^{basic} \epsilon \mathbb{R}^{D \times D^1}$ for feature mapping.
3.   Initialize parameters for batch normalization $\beta, \gamma$ and $\lambda, \alpha$ activation parameters.
4.   For each batch $b$ in $X$:
Apply input Dropout:$X_{drop}^{(b)} = Dropout(X^{(b)})$
Compute attention maps:
For each node i in the tensor:
Compute pairwise node interactions:$Z_{ij}^{(b)} = X_{drop,i}^{(b)} \odot X_{drop,j}^{(b)}$ for $j = 1,2,3 \cdots, L$.
Apply attention transformation: $A_{ij}^{(b)} = \tanh\left(W_{att} Z_{ij}^{(b)} + b_{att}\right)$.
Compute attention coefficients: $\alpha_{ij}^{(b)} = softmax\left(A_{ij}^{(b)}/T\right)$ where $T$ is the temperature
Feature Projection:
For each node $i$:
Aggregate using attention:$Y_{att,i}^{(b)} = \sum_{j=1}^{L} \alpha_{ij}^{(b)} W_{proj}^{att} X_j^{(b)}$
Direct projection:$Y_{basic,i}^{(b)} = W_{proj}^{basic} X_j^{(b)}$
 Combine projections:$Y_i^{(b)} = Y_{att,i}^{(b)} + Y_{basic,i}^{(b)}$
5.   Apply Batch Normalization and Activation
6.   Return the refined feature tensor $Y$ for the Batch
End
```

The Algorithm 1 outlines the operations involving tensor transformations, attention coefficient compotation, and feature refinement, which are critical for emphasizing the most salient features from the input data. The algorithm leverages linear transformations, projection layer and non-linear activations, and normalization techniques to effectively enhance the feature representation for subsequent processing stages in the model. Here, W_{proj}^{basic} refers to projection or feature transformation without attention and W_{proj}^{att} denotes feature projection with attention score.

2.5. Multi-scale feature synthesis and authenticity classification

The proposed module, named multi-scale feature takes the output from the SAG layer and further processes it in a multi-scale manner, potentially applying additional attention mechanisms in a graph-based approach to synthesize features at a higher level of abstraction, capturing both local and global patterns within the feature map $Y_i^{(b)}$, thereby enabling a more robust and comprehensive understanding of the data. The attention mechanism in this module adopts the same function as in SAG but executes in a layered-based manner. From the previous module SAG, the system obtained $Y_i^{(b)}$ which is a result of combining features with attention score $Y_{att,i}^{(b)}$ and without attention score $Y_{basic,i}^{(b)}$. The current module, multi-scale feature synthesis, first initializes two projection layers P1 and P2, which a linear model or simple neural network with layer to obtain two distinct transformations of input feature $Y_i^{(b)}$, such that: $Y1 = P1(Y_i^{(b)})$ and $Y2 = P2(Y_i^{(b)})$. This module also initializes a two-aggregator node A1 and A2, which can also be considered global or aggregate representations influencing the entire graph structure or feature set. This module then uses attention mechanisms (W_{proj}^{att},W_{proj}^{basic}, $W_{proj}^{att\_A1}$, and $W_{proj}^{att\_A2}$) to process Y1 and Y2, considering their relationships with the master nodes. Here, W_{proj}^{att}, and W_{proj}^{basic} apply transformations with and without the influence of the attention maps derived from the standard attention mechanism. While $W_{proj}^{att\_A1}$ and $W_{proj}^{att\_A2}$ are similar but specifically interact with the aggregator node. The aggregator nodes A1 and A2 are updated as A nodes based on the interactions and attention mechanisms, integrating information from Y1, Y2, and the existing state of the master node. After processing through attention and projection layers, the features (and potentially the master node representation) are aggregated. This aggregation is then passed through pooling, dropout, and batch normalization to generate the final output of the layer. Hence, this multi-scale feature synthesis module's primary purpose is to process combined input y, create two separate projections of it (Y1 and Y2), and process these projections through a series of attention mechanisms and linear transformations. The layer integrates these features with aggregator node representations, capturing local (individual feature) and global (aggregate or aggregator node) information. Finally, the authentication classification layer employs a linear model with sigmoid activation to predict the output class bonafide or spoofed. This design allows for a nuanced processing of features, suitable for complex data structures or multi-modal integration tasks in neural networks. Figure 4, illustrates an overall architecture for a proposed system designed to classify audio samples into 'Spoofed' or 'Bonafide' classes of ASV system. The depicted architecture comprises several computing modules, each with a specific role in processing the input data, as discussed in the above sections.

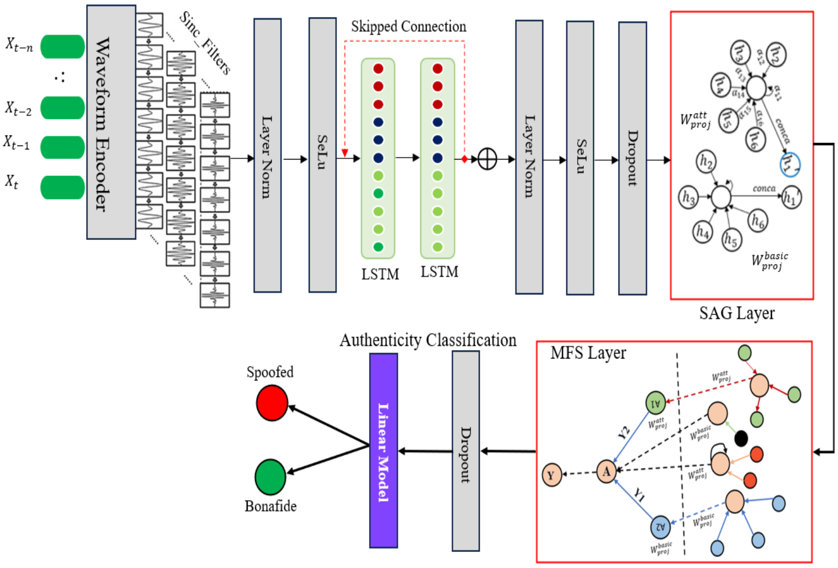

Figure 4. Illustrating overall architecture of the proposed system with deep feature synthesis and selective graph attention mechanism

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 installed on Windows core-i7. The design and training of the proposed model is done with GPU support as it requires substantial computational resources due to sophisticated neural network modules and extensive audio data samples. The performance assessment was strategically focused on the validation dataset due to the substantial size of the validation set, containing more than half the number of samples present in the training dataset. Such a volume of data provides a robust basis for on-the-fly validation, ensuring a comprehensive performance evaluation while optimizing computational resources. Even this strategy of evaluation aligns with methodologies adopted in the wider research community. The study considers different performance metrics for performance evaluation, a loss metric is used to demonstrate training performance, and equal error rate EER () and minimum tandem detection cost function (min t-DCF) both are used to demonstrate model performance of the validation dataset. EER is a metric that offers a standard measure of the trade-off between false acceptance and false rejection rates, while min t-DCF evaluates the system's overall performance by considering both detection effectiveness and the cost of errors. It is particularly relevant for systems where the trade-off between different types of errors is crucial.

Figure 5 presents training loss as a function of epochs. The graph trend shows a decreasing trend in the training loss as the number of epochs increases. This indicates the model's increasing accuracy in predicting the correct classes over time. The initial instability is due to the model exploring the feature space, and the subsequent steady decline in loss suggests that the model is converging towards an optimal set of parameters. The TRU and SAG module's ability to extract and synthesize features at multiple scales plays a crucial role in the model's learning efficiency, as they enable the model to capture both fine-grained and abstract representations of the data, essential for accurate spoofing detection.

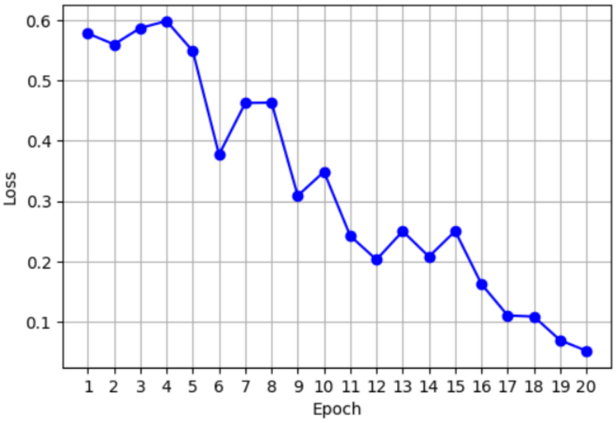

Figure 5. Analysis of training loss over epochs

Figure 6 presents EER analysis over progressive epochs, and the graph trend exhibits EER is highest at the first few epochs, indicating a relatively poor balance between false acceptances and false rejections at the beginning of training. Further, at the 5th epoch, there is a declining trend in EER, indicating a significant improvement in model performance. The model is b ecoming more adept at balancing false acceptances and rejections, enhancing its overall reliability in distinguishing genuine from spoofed samples. The reason behind initial fluctuations could be due to the fact that the model's adjustments to the complexity of the task, while the later stabilization indicates the efficacy of the proposed TRU and SAG modules in refining feature extraction and attention-weight optimization, resulting in a more consistent performance.

Figure 7 shows an analysis of the TDCF metric for evaluating the model's performance in terms of the cost of false positives and false negatives. The graph illustrates that the min t-DCF value is initially high, indicating a greater cost associated with the detection decisions. But after a few epochs decline trend is found, signifying a rapid improvement in the model's detection capabilities and cost-efficiency. Also, some variability in the curve can be seen with a slight increasing and decreasing trend. This suggests that while the model tries to learn optimal features by the SAG module, which leads to recurrent fluctuation when it gets optimal concatenated features in the multi-scale feature synthesis module, it leads to a better and consistent trend. Table 2 presents a comparative analysis to demonstrate the proposed system's effectiveness compared to similar approaches in terms of EER and t-DCF scores.

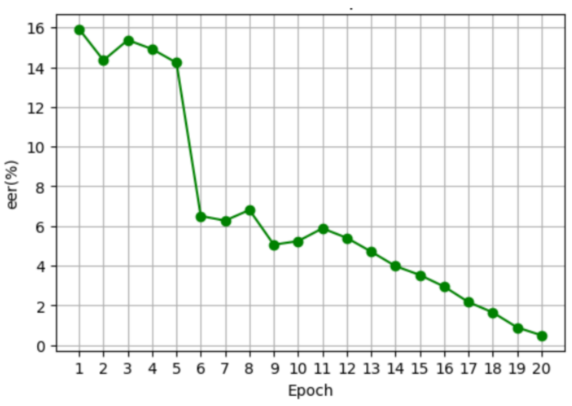

Figure 6. Analysis of EER over progressive epochs

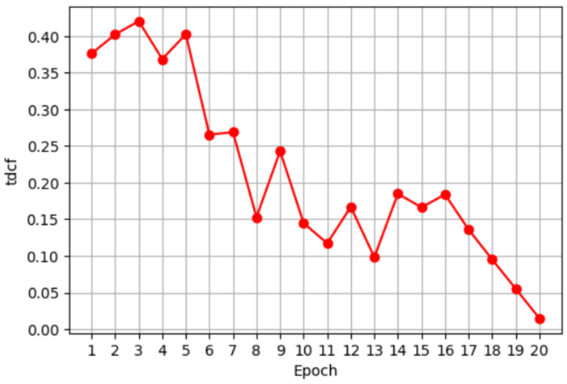

Figure 7. Analysis of t-DCF over progressive epochs

Table 2. Comparative analysis

Methods	EER (%)	t-DCF
Jung *et al*. [29]	0.96	0.0244
Wei *et al*. [30]	1.85	0.0589
Ouyang *et al*. [31]	3.44	0.0800
Proposed	0.503	0.015

Based on the outcome analysis of Table 2, the proposed system outperforms other similar methods. With achieving an EER of 0.503% and a t-DCF of 0.015, it demonstrates its effectiveness in spoofing detection, indicating a more robust detection capability. Jung *et al.* [29] presented a deep learning approach with high-resolution spectrograms, focusing on the direct input of spectrograms without knowledge-based intervention. High-resolution spectrograms can capture finer details and introduce noise or irrelevant information, potentially leading to misclassification. The approach suggested in [30] used acoustic features obtained from linear prediction residual signals and harmonic noise sub-band ratios. These features aim to capture the interaction differences between the vocal tract in genuine and spoofed speech. However, this method struggled with the inherent symmetry between genuine and spoof speech, making it challenging to distinguish between them consistently. While work done by [31] explores the applicability of capsule networks for replay attack detection. Capsule networks have shown effectiveness in forged image and video detection but may not capture the complex temporal relationships present in audio signals, which are crucial for distinguishing between genuine and spoofed speech. On the other hand, the proposed system incorporates a novel integration of waveform encoder and TRU, followed by a selective graph attention mechanism, which offers several advantages over the existing approach, such as robust spectral-temporal feature extraction using waveform encoder and LSTM with skipped connection, graph-based feature processing, attention-based feature importance and extraction of refined feature as an indicative attribute of spoofed signal by multi-scale analysis. Hence, with the advanced feature synthesis process, the proposed system gets superior generalization capabilities that adapt its learning to the evolving nature of spoofing attacks.

4. CONCLUSION

This study has introduced a novel framework for detecting replay attacks on ASV systems. The proposed research work has shown that by integrating spectral-temporal feature extraction with graph-based attention mechanisms, a significant improvement can be achieved in the detection task of sophisticated spoofing attempts. The proposed system basically presented a design of comprehensive and advanced neural network architecture in which the initial module, namely waveform encoder, utilizes SincNet for precise spectral feature extraction integrated with a novel TRU module that consists of LSTM networks with residual connections retaining the temporal dependencies along with spectral information, which is crucial for distinguishing latent cues of spoofing attempts in input audio signal. The study then designed a SAG layer, a joint approach of graph-based feature processing and attention-based feature refinement. The study then strategically implements an additional layer of multi-scale feature synthesis where a graph aggregator node is defined with a multi-scale attention layer to collectively refine and synthesize features at multiple scales, attentively focusing on the most discriminative aspects of the audio signals. The proposed study shows the potential of graph-based architecture and attention-based mechanisms in enhancing the interpretability and focus of specialized neural network models for security-critical applications such as ASV. The results confirm the effectiveness of the proposed system against current state-of-the-art methods concerning both EER and t-DCF metrics. Future work will focus on enhancing the model's scalability and exploring the integration of the reinforcement learning-driven agent model.

REFERENCES

[1] M. Ceaparu, S.-A. Toma, S. Segarceanu, G. Suciu, and I. Gavat, “Multifactor voice-based authentication system,” *Journal of Engineering Science and Technology Review*, pp. 131–136, 2020.

[2] R. C. Johnson, W. J. Scheirer, and T. E. Boult, “Secure voice-based authentication for mobile devices: vaulted voice verification,” in *Biometric and Surveillance Technology for Human and Activity Identification X*, vol. 8712, May 2013, pp. 164-176, SPIE, doi.org/10.48550/arXiv.1212.0042.

[3] 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 & Computer Engineering*, vol. 13, no. 6, pp. 6972-6983, 2023, doi: 10.11591/ijece.v13i6.pp6972-6983.

[4] C. Simon and M. Rajeswari, “Voice-based virtual assistant with security,” in *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, IEEE, 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, 2021, doi: 10.48550/arXiv.2012.00931.

[6] A. Mittal and M. Dua, “Automatic speaker verification systems and spoof detection techniques: review and analysis,” *International Journal of Speech Technology*, pp. 1–30, 2022, doi: 10.1007/s10772-021-09876-2.

[7] 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*, pp. 32725–32762, 2021, doi: 10.1007/s11042-021-11235-x.

[8] 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)*, 2018, doi: 10.23919/APSIPA.2018.8659666.

[9] M. Farrús, “Voice disguise in automatic speaker recognition,” *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–22, 2019, doi: 10.1145/3195832.

[10] 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, 2023, doi: 10.1145/3527153.

[11] 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, 2020, doi: 10.1109/mcom.001.1900396.

[12] 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.

[13] Y. Gong, J. Yang, J. Huber, M. MacKnight, and C. Poellabauer, “ReMASC: Realistic replay attack corpus for voice controlled systems,” *arXiv-Computer Science*, 2019, doi: 10.48550/arXiv.1904.03365.

[14] Z. Wu, S. Gao, E. S. Cling, and H. Li, “A study on replay attack and anti-spoofing for text-dependent speaker verification,” *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Asia-Pacific, Siem Reap, Cambodia, 2014, pp. 1-5, doi: 10.1109/APSIPA.2014.7041636.

[15] L. Li, Y. Chen, D. Wang, and T. F. Zheng, “A study on replay attack and anti-spoofing for automatic speaker verification,” *arXiv-Computer Science*, 2017, doi: 10.48550/arXiv.1706.02101

[16] F. Alegre, A. Janicki, and N. Evans, “Re-assessing the threat of replay spoofing attacks against automatic speaker verification,*” 2014 International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, 2014, pp. 1-6.

[17] J. Xue *et al*., “Learning from yourself: A self-distillation method for fake speech detection,” *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096837.

[18] Z. Lei, Y. Yang, C. Liu, and J. Ye, “*Siamese Convolutional Neural Network Using Gaussian Probability Feature for Spoofing Speech Detection*,” in *Interspeech*, 2020, pp. 1116–1120, doi: 10.21437/Interspeech.2020-2723.

[19] Z. Wu, R. K. Das, J. Yang, and H. Li, “Light convolutional neural network with feature genuinization for detection of synthetic speech attacks,” *arXiv-Electrical Engineering and Systems Science*, 2020, doi: 10.48550/arXiv.2009.0963

[20] A. Javed, K. M. Malik, H. Malik, and A. Irtaza, “Voice spoofing detector: A unified anti-spoofing framework,” *Expert Systems with Applications*, vol. 198, 2022, doi: 10.1016/j.eswa.2022.116770.

[21] 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.

[22] 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*, 2022, doi: 10.1145/3552466.3556533.

[23] J. Guo, Y. Zhao, and H. Wang, “Generalized spoof detection and incremental algorithm recognition for voice spoofing,” *Applied Sciences*, vol. 13, no. 13, 2023, doi: 10.3390/app13137773.

[24] M. S. Saranya, R. Padmanabhan, and H. A. Murthy, “Replay attack detection in speaker verification using non-voiced segments and decision level feature switching,” in *2018 International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, India, 2018, pp. 332-336, doi: 10.1109/SPCOM.2018.8724469.

[25] P. J. Kemanth, S. Supanekar, and S. G. Koolagudi, “Audio replay attack detection for speaker verification system using convolutional neural networks,” in *Pattern Recognition and Machine Intelligence: 8th International Conference, PReMI 2019, Tezpur, India*, 2019, pp. 445-453, Springer International Publishing.

[26] S. -H. Yoon, M. -S. Koh, J. -H. Park, and H. -J. Yu, “A new replay attack against automatic speaker verification systems,” in *IEEE Access*, vol. 8, pp. 36080-36088, 2020, doi: 10.1109/ACCESS.2020.2974290.

[27] A. Nautsch *et al*., “ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech,” in *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252-265, April 2021, doi: 10.1109/TBIOM.2021.3059479.

[28] M. Ravanelli and Y. Bengio, “Speech and speaker recognition from raw waveform with sincnet,” *arXiv-Electrical Engineering and Systems Science*, 2018, doi: 10.48550/arXiv.1812.05920

[29] J. W. Jung, H. J. Shim, H. S. Heo, and H. J. Yu, “Replay attack detection with complementary high-resolution information using end-to-end dnn for the asvspoof 2019 challenge,” *arXiv-Electrical Engineering and Systems Science*, 2019, doi: 10.48550/arXiv.1904.10134

[30] L. Wei, Y. Long, H. Wei, and Y. Li, “New acoustic features for synthetic and replay spoofing attack detection,” *Symmetry*, vol. 14, no. 2, 2022, doi.org/10.3390/sym14020274.

[31] M. Ouyang, R. K. Das, J. Yang, and H. Li, “Capsule network based end-to-end system for detection of replay attacks,” *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Hong Kong, 2021, pp. 1-5, doi: 10.1109/ISCSLP49672.2021.9362111.

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 a 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.