scientific reports

Check for updates

OPEN

DeepLASD countermeasure for logical access audio spoofing

Hamed Al-Tairi[1], Ali Javed[2], Tasawer Khan[3] & Abdul Khader Jilani Saudagar[4]✉

Voice-based authentication systems have become increasingly vulnerable to logical access (LA) spoofing through sophisticated voice conversion (VC) and text-to-speech (TTS) attacks. This paper proposes an end-to-end deep learning approach DeepLASD, that processes raw waveforms to detect spoofed speech without relying on handcrafted features. The model incorporates a SincConv layer for interpretable spectral processing, along with residual convolutional blocks that integrate attention for improved feature extraction. We introduce GeLU activation in residual blocks to enhance our method's ability to better capture the unique traits in real and spoof samples. A gated recurrent unit is further employed for temporal dynamics modeling. Extensive experimentation was conducted on the large-scale and diverse ASVspoof 2019 and 2021 datasets. Achieving an Equal Error Rate as low as 4.98% and a minimum Tandem Detection Cost Function of 0.1208, along with strong generalization to both VC and TTS spoof types, demonstrate the competency of the proposed method for LA spoofing detection. Although the results on the ASVspoof 2021 dataset underscore the challenges posed by next-generation synthetic speech, the proposed solution exhibits notable adaptability. These findings affirm that the proposed end-to-end anti-spoofing framework enhances security and detection capabilities in voice authentication systems.

Keywords Automatic speaker verification, Deep learning, Logical access attacks, Spoof detection, Text-to-speech synthesis, Voice conversion

Automatic Speaker Verification (ASV) has emerged as a convenient and secure method of authenticating users based on vocal characteristics. Its applications span industries such as finance, healthcare and the Internet of Things (IoT), enabling hands-free voice-based identity verification. Despite its advantages, ASV systems face a growing range of spoofing attacks that pose serious security concerns[1,2]. These attacks aim to deceive ASV systems through methods such as replaying pre-recorded speech, employing advanced voice conversion (VC), generating text-to-speech (TTS) audio, or even creating imperceptibly altered inputs using adversarial techniques.

Among the various spoofing attacks, logical access (LA) attacks have gained special attention due to the increasing sophistication of synthetic speech technologies[3]. Rather than interacting with an ASV system via the typical acoustic input path (e.g., a microphone), logical access attacks feed digitally generated signals directly into the software chain, effectively bypassing hardware-based controls. Advanced TTS and VC methods can craft high-fidelity speech samples, making detection considerably more challenging and increasing the risk of fraudulent activity. Consequently, enhancing ASV systems to detect these attacks is vital for ensuring their reliability and trustworthiness.

The significance of this study is threefold. First, it contributes to a deeper understanding of vulnerabilities within ASV systems and highlights the urgent need for more adaptive and generalized countermeasures[1]. Second, by demonstrating robust performance on the ASVspoof 2019 and 2021 datasets, it underscores the importance of leveraging large-scale community-driven benchmarks. Finally, as voice-driven interfaces become more prevalent, ensuring the security and reliability of ASV technology has profound legal, cultural, and ethical implications. Tackling logical access spoofing effectively will help maintain user trust, protect sensitive information, and meet emerging regulatory standards for biometric authentication systems.

Existing anti-spoofing methods frequently rely on handcrafted spectral features, which can limit their ability to generalize to advanced or unseen spoofing techniques. Moreover, many approaches lack robust mechanisms for capturing the subtle spectral and temporal artifacts introduced by sophisticated VC and TTS methods. These limitations often lead to degraded performance in cross-dataset evaluations or emerging attack scenarios.

[1]School of Information Technology, Whitecliffe College, Auckland, New Zealand. [2]Department of Software Engineering, University of Engineering and Technology, Taxila, Pakistan. [3]James Watt School of Engineering, University of Glasgow, Glasgow, UK. [4]Information Systems Department,College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), 11432 Riyadh, Saudi Arabia. ✉email: aksaudagar@imamu.edu.sa

To address these challenges, there is a need to develop a robust end-to-end deep learning framework that directly processes raw audio signals, eliminating the need for handcrafted features and adapting more effectively to evolving spoofing techniques. We leveraged robust activation functions (e.g., GeLU, LeakyReLU) in combination with SincNet-based convolution and attention mechanisms to capture subtle frequency distortions. We also explored key hyperparameter and architectural configurations that further enhance spoof-detection performance. While prior end-to-end systems such as RawNet and RawNet2 have demonstrated the potential of raw waveform processing for spoofing detection, this work focuses on further improving the existing RawNet and RawNet2 models by optimizing the integration of interpretable SincConv filters, attention mechanisms, and advanced activation functions to enhance adaptability and generalization across diverse spoofing attacks. Rather than introducing entirely new architectural components, we contribute a carefully tuned combination of proven modules and provide comprehensive empirical validation, enhancing system performance on challenging logical access tasks.

The main contributions of this study are summarized as follows:

- We propose an end-to-end DL method DeepLASD centered on SincNet-based convolution and attention mechanism for effective LA spoofing detection.
- We introduce GeLU activation in residual blocks that makes our method learn salient cues from the audio and stabilize training for improved performance.
- We performed comprehensive experiments on ASVspoof 2019 and 2021 datasets, including ablation study, to demonstrate robust generalization of our method to TTS and VC attacks.

Related work

Research on ASV systems has progressively revealed critical security weaknesses that malicious actors exploit to gain unauthorized access. Broadly, these attacks can be categorized into physical-access (PA), logical-access (LA), and adversarial attacks[4,5]. Physical attacks, such as replaying pre-recorded speech or impersonating a legitimate user, pose an immediate threat by injecting counterfeit audio directly through microphones. Logical-access attacks bypass hardware sensors altogether by feeding manipulated or synthesized speech into the software pipeline, while adversarial attacks introduce barely perceptible distortions that can drastically degrade ASV system accuracy[6].

Early attempts at mitigating these threats relied on handcrafted features, including Mel-frequency cepstral coefficients (MFCCs), linear frequency cepstral coefficients (LFCCs), and constant Q cepstral coefficients (CQCCs)[7–9]. These features were combined with traditional machine learning models, such as Gaussian mixture models (GMMs) and support vector machines (SVMs). Although they proved effective for straightforward replay or voice conversion scenarios, their adaptability to higher fidelity or advanced attacks remained limited[10].

As computational power increased, researchers leveraged deep learning to enhance spoof detection, typically by blending neural networks with existing handcrafted feature sets[11,12]. For instance, lightweight convolutional neural network (CNN) architectures improved detection efficiency and reduced latency[13], while more advanced back-end classifiers tackled complex replay conditions. However, many of these solutions still depended on engineered descriptors, constraining their flexibility in responding to evolving spoof tactics.

In an effort to eliminate reliance on manual feature extraction, end-to-end frameworks began to emerge. Models like RawNet[14,15] and others process raw audio waveforms directly, capturing subtle spectral and phase cues missed by handcrafted features. Advanced end-to-end models such as Stat-SE-Res2Net50, which integrate squeeze-and-excitation modules with residual blocks, have also demonstrated strong performance in spoof detection tasks[16]. These methods achieved significant improvements across various spoof categories, including voice conversion and text-to-speech synthesis, thereby underscoring the potential of raw-audio pipelines for handling increasingly sophisticated attacks[17,18].

Simultaneously, advancements in speech synthesis and generative modeling enabled attackers to produce near-human-quality voices[19,20]. Logical-access spoofing thus expanded in scope, including speech generated via deep neural networks or adversarial perturbations[21,22]. Traditional countermeasures often struggled to generalize to these high-fidelity spoofs, prompting researchers to explore deeper architectures (ResNet, DenseNet) and attention mechanisms that focus on critical time-frequency regions[23,24].

Building upon these innovations, some studies pursued *unified solutions* capable of detecting replay, synthetic speech, and cloned voices within a single framework[12,25]. By covering multiple spoof types concurrently, unified approaches can more holistically protect ASV systems under real-world conditions where attacks can vary widely. However, designing a one-size-fits-all architecture remains complex, as different spoofing methods exhibit diverse artifacts that require specialized feature representation[1].

Another line of work embraced *integrated solutions*, wherein spoof detection and ASV are combined into a single pipeline[26,27]. Such integrated methods attempt to verify the speaker and assess authenticity in tandem, reducing system overhead and potentially improving security. Nonetheless, unifying these two objectives demands carefully balanced multi-task learning or joint optimization, and many solutions are still tested under limited conditions, making it difficult to ascertain their robustness[28].

In parallel with algorithmic advancements, the research community introduced new datasets and benchmarks to promote rigorous evaluation. The ASVspoof series (2019, 2021)[3,29] incorporates both logical and physical access scenarios, as well as different forms of speech synthesis. Additional corpora, such as the Voice Spoofing Detection Corpus (VSDC), expand replay conditions with multi-order configurations[30]. These resources facilitate standardized cross-dataset comparisons and highlight shortcomings in existing methods when faced with novel or high-complexity spoofs.

Assessment metrics have also evolved. While Equal Error Rate (EER) remains a core indicator of system accuracy, the tandem Detection Cost Function (t-DCF)[31] provides a cost-sensitive perspective that balances

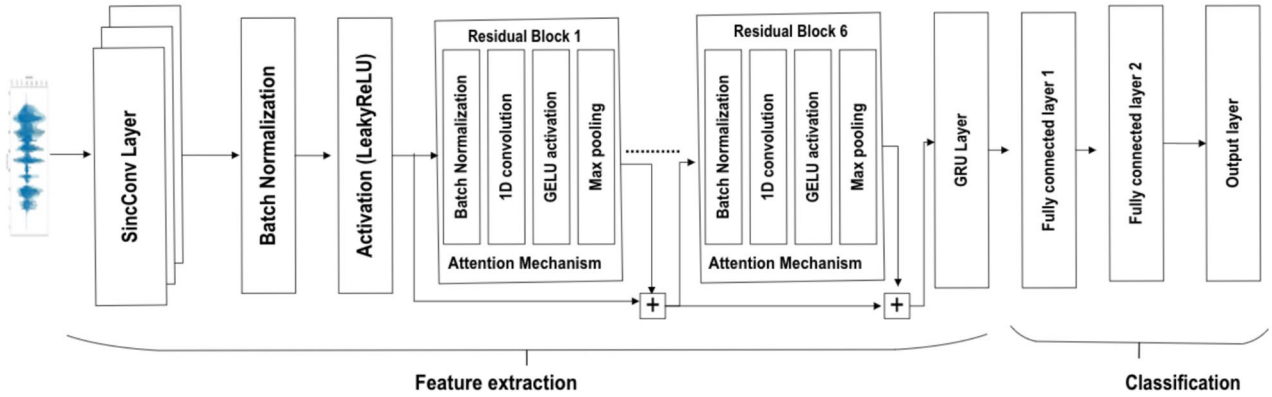

Fig. 1. Proposed End-to-End DeepLASD architecture.

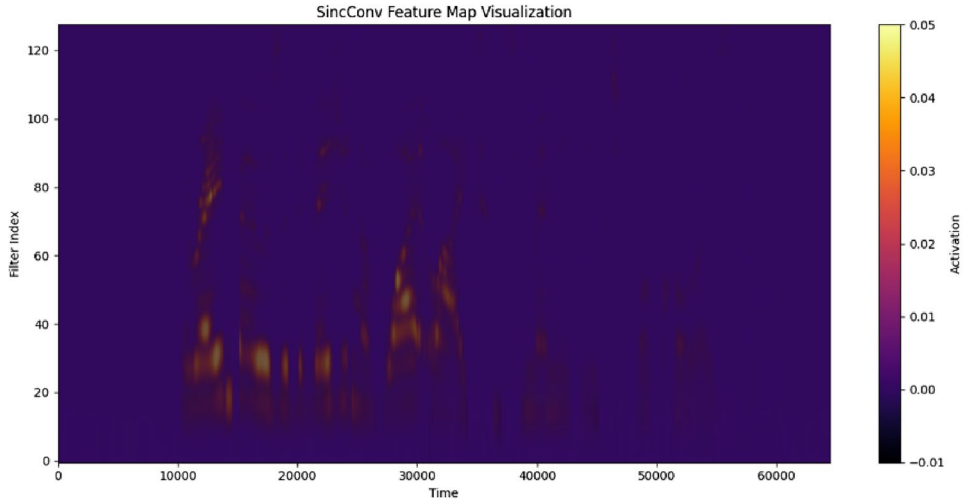

Fig. 2. SincConv feature map visualization.

missed detections and false alarms under realistic operational assumptions. The ASVspoof challenges have consistently used EER and min-tDCF to facilitate fair, application-driven evaluations of spoofing countermeasures.

Despite these concerted efforts, current gaps persist in addressing sophisticated logical-access spoofing, particularly in unseen conditions or rapidly evolving neural synthesis approaches[1]. Models often fail to generalize to complex VC or TTS outputs, and adversarially crafted inputs can further undermine performance. Moreover, though advanced activation functions and attention-based modules offer promise, selecting the ideal architecture for a given spoof scenario remains non-trivial. Our proposed work, therefore, advances beyond prior solutions by integrating SincNet-based convolution layers, embedding of GeLU activation in residual blocks, and attention mechanism to reliably capture salient cues and frequency distortions of modern LA attacks with improved adaptability.

Methodology

The framework proposed in Figure 1 addresses logical access spoofing attacks, specifically those based on voice conversion and text-to-speech synthesis, by processing raw audio signals in an end-to-end manner, eliminating the need for handcrafted features. Instead, the model learns spoof-specific representations directly from waveforms, enabling flexible adaptation to emerging attack methods. The overall architecture is composed of four primary components: a SincConv layer, a stack of residual convolution blocks with integrated attention, a Gated Recurrent Unit (GRU) layer, and a final classification head. An overview of the architecture is presented in Figure 1.

Proposed end-to-end DeepLASD architecture

The architecture is specifically designed to capture algorithmic artifacts of spoof samples and dynamic variations of bonafide audios for better LA spoofing detection. Each component contributes uniquely to the overall performance, as detailed below.

SincConv layer (Component A)
At the network's front-end, the SincConv layer replaces conventional convolutional filters with parameterized sinc functions, providing a more interpretable and efficient frequency analysis. Speech signals, $x(t)$, are decomposed into distinct frequency bands using band-pass filters defined by learnable cutoff frequencies f_1 and f_2. The impulse response of each filter is given by:

$$g(t; f_1, f_2) = 2f_2 \,\mathrm{sinc}(2\pi f_2 t) - 2f_1 \,\mathrm{sinc}(2\pi f_1 t) \tag{1}$$

where the sinc function is defined as:

$$\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x} \tag{2}$$

The convolution operation for a given input signal $x(t)$ and the filter $g(t)$ is expressed as:

$$y(t) = x(t) * g(t) = \int_{-\infty}^{\infty} x(\tau)\, g(t-\tau; f_1, f_2)\, d\tau \tag{3}$$

The convolution operation produces feature maps that highlight frequency bands where spoofing artifacts occur. Key hyperparameters—filter count, length, and regularization on cutoff frequencies—balance temporal resolution and frequency discrimination. Our SincConv layer replaces traditional filters with parameterized sinc functions for interpretable frequency decomposition. To prevent overfitting, we initialize filters using a Mel-scale filterbank (with experiments on linear and inverse-Mel scales) and fix the raw waveform to 4 seconds (64000 samples). We also optimized the filter length and found that 129 samples yields superior performance. Figure 2 illustrates how different SincConv filters activate over time for a sample input. This visualization highlights the layer's ability to extract meaningful, frequency-specific features directly from raw waveforms, supporting its role in detecting spoofed audio.

Residual convolution blocks with attention (Component B)
Following the SincConv layer, six residual convolution blocks extract multi-scale features, with independent feature map scaling highlighting the most informative outputs. Each block is mathematically represented by a residual mapping:

$$\mathrm{y} = \mathrm{x} + \mathcal{F}(\mathrm{x}, \{W_i\}) \tag{4}$$

where x is the input to the block, $\mathcal{F}(\cdot)$ represents the transformation (convolution, non-linear activation, and normalization) with weights $\{W_i\}$, and the sum denotes the residual (or shortcut) connection.

Attention Mechanisms Integrated attention mechanisms refine the feature maps by dynamically weighting the spatial and channel dimensions. A simple formulation for spatial attention is:

$$\mathrm{A}_{\mathrm{spatial}} = \sigma\left(W_{\mathrm{spatial}} * \mathrm{F}\right) \tag{5}$$

where F is the feature map, W_{spatial} is a learnable weight filter, $*$ denotes convolution, and $\sigma(\cdot)$ is a sigmoid activation function that normalizes the attention weights between 0 and 1.

Channel attention can be similarly expressed as:

$$\mathrm{A}_{\mathrm{channel}} = \sigma\left(W_{\mathrm{channel}} \cdot \mathrm{GAP}(\mathrm{F})\right) \tag{6}$$

where $\mathrm{GAP}(\cdot)$ denotes Global Average Pooling applied over spatial dimensions, and W_{channel} is a learnable weight matrix.

The output of the attention module is then combined with the original feature map:

$$\mathrm{F}_{\mathrm{att}} = \mathrm{F} \odot (\mathrm{A}_{\mathrm{spatial}} + \mathrm{A}_{\mathrm{channel}}) \tag{7}$$

where $\odot$ represents element-wise multiplication.

We introduce GeLU activation in the convolutions of residual blocks followed by the attention mechanism to capture complex traits in the audio. Batch normalization or layer normalization is also applied to stabilize training.

GRU layer
To model the temporal dynamics of speech, a GRU layer is employed. The GRU updates its hidden state using the following equations:

$$z_t = \sigma\left(W_z x_t + U_z h_{t-1} + b_z\right) \tag{8}$$

$$r_t = \sigma\left(W_r x_t + U_r h_{t-1} + b_r\right) \tag{9}$$

$$\tilde{h}_t = \tanh\left(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h\right) \tag{10}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{11}$$

where x_t is the input at time t, h_t is the hidden state, z_t is the update gate, r_t is the reset gate, and $\odot$ denotes element-wise multiplication. The weights $W_{\{z,r,h\}}$, $U_{\{z,r,h\}}$, and biases $b_{\{z,r,h\}}$ are learnable parameters. The GRU layer is added to integrate frame level embeddings into a single utterance level embedding. The GRU layer efficiently captures both short-term and long-term dependencies in the sequential data.

Classification head
The final stage of the architecture is the classification head, which aggregates the learned features and produces a binary decision. Let h denote the aggregated feature vector from the GRU layer. A series of fully connected (dense) layers transforms h:

$$\mathrm{z} = \phi\left(W_{\mathrm{fc}}\mathrm{h} + b_{\mathrm{fc}}\right) \tag{12}$$

where W_{fc} and b_{fc} are the weights and biases of the dense layer, and $\phi(\cdot)$ is a non-linear activation function. Finally, the softmax function converts the output z into class probabilities:

$$p_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \tag{13}$$

where p_i is the probability of class i (genuine or spoofed). The network is trained using a cross-entropy loss function, defined as:

$$\mathcal{L} = -\sum_i y_i \log(p_i) \tag{14}$$

with y_i representing the ground-truth label.

Activation function configuration

The activation function computes the output of nodes in neural network according to its inputs and corresponding weights. We employed GeLU activation in residual blocks and Leaky ReLU in Sinc convolution block. We selected GeLU in residual blocks because of its smooth gradient and nonzero gradient for negative inputs. These attributes improve the efficacy of residual connections by making the architecture learn complex traits while keeping stable training. More precisely, GeLU is smooth and differentiable, with probabilistic interpretation, better capability to handle negative values and regularization effect. These attributes make GeLU a reliable activation function in residual blocks of proposed architecture to achieve better performance and more stable training.

Similarly, the Leaky Rectified Linear Unit (Leaky ReLU) is utilized in the Sinc convolution block to address the limitations inherent in standard ReLU functions. Leaky ReLU allows a small, nonzero gradient when the input is negative, which prevents neurons from becoming inactive, a common issue with standard ReLU known as the "dying ReLU" problem. This characteristic ensures continuous gradient flow during training, thereby facilitating the capture of subtle non-linear features essential for effective spoofing detection. The efficiency and simplicity of Leaky ReLU make it a robust alternative in scenarios where nuanced feature extraction is critical.

Although the standard Rectified Linear Unit (ReLU) is widely used due to its computational efficiency, its tendency to produce "dead neurons" where negative inputs result in zero output limits its effectiveness in learning intricate features. For this reason, standard ReLU is used sparingly in our architecture, with advanced alternatives such as GeLU, Leaky ReLU, and SELU being preferred for their superior handling of gradient flow and learning dynamics.

This careful selection and assignment of activation functions ensure that the network adapts robustly to a variety of synthetic speech distortions while maintaining effective learning dynamics.

Training procedure and hyperparameters

Training is conducted over 100 epochs using the Adam optimizer with an initial learning rate of 0.0001 and a weight decay of 0.0001. A batch size of 32 is selected to balance computational efficiency and model stability. To further enhance training:

- A learning rate scheduler reduces the learning rate when performance plateaus, facilitating continued improvement in model accuracy.
- Early stopping is implemented based on validation performance, preventing overfitting by halting training once no further improvements are observed.

The proposed method presents an effective end-to-end DL approach for logical access spoofing detection by processing raw audios. From the SincConv layer's frequency-specific filtering (Eqs. 1 and 3) to the multi-scale feature extraction in the residual convolution blocks with attention (Eqs. 4–7), the GRU's temporal modeling (Eqs. 8–11), and the finely tuned classification head (Eqs. 12–13), each element is meticulously engineered to meet the challenges posed by synthetic speech detection.

Experiments and Results

This section presents the experimental setup and performance outcomes of the proposed End-To-End DeepLASD Detector on both the ASVspoof 2019[32] and ASVspoof 2021[33] Logical Access datasets. We begin with a description of data preparation and the evaluation metrics used, followed by a series of paragraphs summarizing key experimental findings.

Spoof type	Dataset	min-tDCF	EER (%)
Overall LA-eval	ASVspoof 2019-LA-Eval	0.1216	5.2753
Text-to-speech	ASVspoof 2019-TTS	0.1442	6.6891
Voice conversion	ASVspoof 2019-VC	0.1477	6.2678
Overall LA-eval	ASVspoof 2021-LA-Eval	0.4250	12.76

Table 1. Performance evaluation of the proposed DeepLASD method.

Exp #	Configuration	EER (%)	Min-tDCF
1	ReLU (residual block) + ReLU (Sinc block)	7.95	0.1413
2	LeakyReLU (residual block) + SELU (Sinc block)	4.98	0.1208
3	GELU (residual block) + LeakyReLU (Sinc block)	5.28	0.1216

Table 2. Comparative analysis of activation functions on ASVspoof 2019 LA-eval Dataset.

We employed the ASVspoof 2019[32] and ASVspoof 2021[33] datasets as primary benchmarks, each featuring a variety of spoofing scenarios, including VC and TTS attacks[3,29]. Each dataset includes predefined training, development, and evaluation subsets, which we used for model training, validation, and final testing respectively, following ASVspoof challenge protocols. Where necessary, audio files were resampled to 16 kHz, normalized to a fixed amplitude range, and segmented to ensure consistent input dimensions for the detector. The ASVspoof 2019 dataset served as the main training and development resource, while ASVspoof 2021 allowed us to evaluate generalization to more challenging spoofing techniques.

We reported performance using two key metrics recommended in ASVspoof challenges. The first is the Equal Error Rate (EER), defined by the operating point at which false acceptance and false rejection rates are equal. The second is the Minimum Tandem Detection Cost Function (min-tDCF)[31], a cost-sensitive measure incorporating misclassification costs and prior probabilities. Lower EER and min-tDCF values denote improved spoof detection, and both are evaluated on the respective 2019 and 2021 LA evaluation subsets[3,29].

Performance Evaluation of proposed method

We designed an in-depth multi-stage experiment for assessing the effectiveness of proposed antispoofing detector. For this purpose, we evaluated our DeepLASD model on the ASVspoof 2019 and 2021 LA datasets overall, and TTS and VC subsets separately. We used the same hyperparameter settings as mentioned in Section III-C. We used the train set for model training, and eval set for model testing for all experiments. Our detector achieved a min-tDCF of 0.1216 and an EER of 5.2753% on ASVspoof 2019-LA-Eval. For the TTS and VC subsets, the min-tDCF/EER pairs were 0.1442/6.6891% and 0.1477/6.2678%, respectively, while on ASVspoof 2021-LA-Eval, we observed a slight decline in performance by attaining a min-tDCF of 0.4250 and EER of 12.76%, reflecting the increased complexity of ASVspoof 2021 dataset. Still, we achieved competitive performance for LA spoofing detection. These results in Table 1 underscore the ability of our method to capture subtle spoof artifacts, attributed to the use of GELU in residual blocks and LeakyReLU in the Sinc convolution block for stable gradient flow and nuanced feature extraction.

Ablation study

This experiment was designed to examine the effects of different architectural configurations in our DeepLASD method. We investigated how advanced activation functions and specific architectural configurations affect performance on both ASVspoof 2019 and 2021 LA datasets. The experiments were carried out using the same experimental and hyperparameter settings as mentioned above, with results highlighting the sensitivity of our method to different design choices.

A. Impact of different activation functions. This study was performed to investigate the effect of different activation functions in residual and sinc convolution blocks in our DeepLASD. We conducted more than 20 experiments exploring a range of activations (e.g., ReLU, LeakyReLU, SELU, GELU) in our method to determine the configuration, which achieves the lowest EER and min-tDCF values. We reported the details of top three performing activation configurations in Table II. From the results in Table 2, we conclude that pairing GELU in residual layers with LeakyReLU in the Sinc convolution layer consistently yielded the lowest EER (5.28%) and a stable training process. This study highlights the efficacy of GeLU activation in residual blocks and leaky ReLU in Sinc convolution blocks for better feature computation of spoofing artifacts in synthetic samples and speech dynamics of bonafide samples.

B. Investigation of different strategies of Pooling and FC layers. This two-stage experiment was planned to examine the effect of different pooling strategies and impact of adding an additional FC layer. First, we carried out this experiment by examining the effect of using average pooling instead of max-pooling and results are shown in Table III. Next, we introduced an additional FC layer in addition of two FC layers to assess its impact on the spoofing detection performance. As shown in Table 3, adding a third FC layer led to a modest increase in EER (from 5.28% to 6.39% on ASVspoof 2019), while substituting max pooling with average

Dataset	Configuration	Min-tDCF	EER (%)
2019 LA-Eval	Proposed DeepLASD (2 FC, max pool, GELU+LeakyReLU)	0.1216	5.28
2019 LA-Eval	DeepLASD (3 FC layers, max pool, GELU+LeakyReLU)	0.1396	6.39
2019 LA-Eval	DeepLASD (2 FC layers, Avg pool, GELU+LeakyReLU)	1.0000	48.18
2021 LA-Eval	DeepLASD (3 FC layers, max pool, GELU+LeakyReLU)	0.9999	47.10

Table 3. Comparative analysis of different pooling strategies and FC Layers on ASVspoof 2019 and 2021 LA Datasets.

Methods	min-tDCF	EER (%)
CQCC-GMM baseline[34]	0.236	9.87
LFCC-GMM baseline[34]	0.212	11.96
FBCC-GMM[35]	0.155	6.16
Stat-SE-Res2Net50[16]	0.068	2.86
LFCC+ProdSpec+MGDCC-CNN[36]	0.198	9.09
CQT+LFCC+DCT-LCNN[37]	0.051	1.84
ATCoP+GTCC-SVM[38]	0.050	0.75
RawNet2-S1[39]	0.1301	5.64
RawNet2-S2[39]	0.1175	5.13
RawNet2-S3[39]	0.1294	4.66
L+S1[39]	0.0330	1.12
L+S1+S2+S3[39]	0.0347	1.14
L+S3[39]	0.0370	1.14
Proposed DeepLASD method	**0.1216**	**5.2753**

Table 4. Comparison of proposed and baseline methods on ASVspoof 2019-LA-Eval.

pooling dramatically worsened performance (EER rising to 48.18%). Similar trends on the ASVspoof 2021 subset confirm that maintaining a balanced architecture with max pooling and controlled network depth using two FC layers is key to effectively detecting spoofing.

These ablation studies reveal that the combination of GeLU and ReLU in residual and sinc convolution blocks, respectively, and configurations involving two FC layers and max pooling for downsampling ensure the best generalization, particularly for confronting more sophisticated, unseen spoofing attacks.

Comparative analysis with baseline and SOTA methods

We conducted a series of around five additional experiments to compare the performance of proposed DeepLASD method against widely known baselines (e.g., CQCC-GMM, LFCC-GMM) and state-of-the-art systems on the ASVspoof 2019 LA dataset. Trained over 100 epochs on the LA partition, our model was evaluated under various spoofing scenarios (text-to-speech, voice conversion). Table 4 highlights that our DL approach, powered by embedding selected activations, yields an EER of 5.2753% and a min-tDCF of 0.1216, outperforming or closely matching several established techniques. Notably, RawNet2-S2 achieves a slightly better min-tDCF (0.1175) but at a similar EER (5.13%), while RawNet2-S1 and S3 show comparable performance ranges. These comparisons help illustrate the trade-offs between our interpretable modular design and fully end-to-end raw waveform pipelines.Despite some hand-crafted feature methods that excel in niche conditions, these results underscore the adaptability and effectiveness of a direct waveform-based pipeline in detecting subtle spoof artifacts across multiple attack types.

Comparative analysis on voice conversion

Voice conversion spoofing is more difficult to detect than TTS spoofing. To assess the robustness of our DeepLASD for voice conversion, we performed a set of focused trials on the ASVspoof 2019 VC subset to determine if our End-to-End DeepLASD surpassed established baseline methods. Again, we used the same experimentation protocols as mentioned in Section III-C. As shown in Table 5, our method achieved a min-tDCF of 0.1477 and an EER of 6.2678%, outperforming all baseline approaches by a wide margin. These results demonstrate the efficacy of the proposed DeepLASD for capturing intricate frequency cues that feature-based solutions often miss. We can conclude from this experiment that our DeepLASD method can reliably be employed to detect the challenging VC spoofing.

Methods	min-tDCF	EER (%)
Baseline CQCC[40]	0.339	30.61
Baseline LFCC[40]	0.714	46.50
Baseline ELTP-LFCC[40]	0.390	33.28
Proposed (GELU + LeakyReLU, slope=0.3)	**0.1477**	**6.2678**

Table 5. Comparison of proposed and baseline methods for voice conversion detection.

Theoretical rationale

The proposed DeepLASD design is grounded in key principles from signal processing and deep learning. The SincConv layer provides interpretable, learnable bandpass filtering for raw waveforms [41], offering a principled alternative to traditional CNN kernels. GeLU activation [42] ensures smooth, stable gradient flow, enhancing training compared to standard ReLU. Attention mechanisms [43] enable the model to dynamically prioritize the most informative temporal and spectral features, effectively capturing spoofing artifacts. Together, these components create a theoretically informed architecture tailored for logical access spoofing detection.

Discussion

Our experimental results collectively underscore the effectiveness of our end-to-end, DeepLASD for logical access spoof detection, offering several notable insights. First, direct waveform processing with a SincConv-based front end reduces dependency on handcrafted descriptors (e.g., LFCC, CQCC) and adapts more effectively to advanced voice conversion and text-to-speech scenarios. Second, carefully selected activation functions, particularly GELU in residual blocks paired with LeakyReLU in the Sinc convolution block, consistently produced a lower EER and min-tDCF. This pairing alleviates vanishing gradients and captures subtle spectral-temporal cues crucial for distinguishing genuine audio from spoofed samples.

Benchmark comparisons show that the proposed model remains highly competitive with established baselines and state-of-the-art systems. Our method outperformed most handcrafted features-based methods and the two RawNet versions. Although a few fusion-based methods comprising handcrafted features and deep learning approaches slightly outperform DeepLASD, but, at the expense of increased computational cost. Its resilience extends to a variety of spoofing types, including TTS and VC. In more detailed VC-focused testing, the architecture significantly outperformed feature-engineered methods, demonstrating that raw waveform analysis can effectively detect higher-frequency or phase-related artifacts absent in many handcrafted features. Furthermore, experiments with architectural variations highlight the importance of balanced design choices: adding extra layers or substituting max pooling with average pooling often leads to overfitting or performance degradation, especially on more complex datasets like ASVspoof 2021. These observations point to a careful trade-off between network depth, pooling strategies, and computational overhead when deploying in practical settings. Although some baselines outperform in min-tDCF or EER, they often rely on larger models or specialized handcrafted features. In contrast, DeepLASD offers complementary strengths by using a lightweight end-to-end design that learns directly from raw waveforms, enhancing generalization and practical deployment. Although some baselines outperform in min-tDCF or EER, they often rely on larger models or specialized handcrafted features. In contrast, DeepLASD offers complementary strengths by using an interpretable, lightweight end-to-end design that learns directly from raw waveforms, enhancing generalization and practical deployment.

Taken together, the evidence suggests that end-to-end deep learning, combined with attention mechanisms and advanced activations, can reliably detect even sophisticated logical-access spoofing attempts. However, more challenging conditions in newer datasets (e.g., ASVspoof 2021) imply the need for domain adaptation and adversarial strategies to handle rapidly evolving generative models. Additionally, focusing on lightweight, efficient designs and robust training schemes (e.g., data augmentation, semi-supervised learning) will be essential for real-world deployment, particularly where computational resources and latency are constrained.

It is important to emphasize that the novelty of our work lies in the innovative architectural composition of established modules (e.g., SincConv, attention, GRU, GeLU), which are integrated and orchestrated in a unified framework tailored specifically for robust deepfake detection. Rather than introducing new components, we contribute a novel configuration that leverages the complementary strengths of these elements. Our design is driven by a principled understanding of multimodal dynamics and is empirically validated to consistently perform well across a broad spectrum of spoofing attacks. This demonstrates that architectural innovation can emerge from strategic design and synergy, not solely from inventing new components.

Conclusions

This work has presented an end-to-end deep learning approach for logical access spoof detection in ASV systems. By removing the reliance on handcrafted features, the architecture adapts smoothly to various spoofing forms, showing strong performance on the diverse and largescale ASVspoof 2019 and 2021 datasets and maintaining competitive metrics relative to state-of-the-art systems. Our ablation study emphasized the importance of embedding GeLU and LeakyReLU activations in convolution layers to capture salient cues and achieve stable training. While the framework proved robust to a range of attacks, performance on the ASVspoof 2021 dataset revealed challenges introduced by increasingly realistic and complex synthetic speech. Future work may include adopting adversarial training techniques, exploring domain adaptation strategies, and refining architectural elements to enhance generalization across emerging spoof scenarios. Overall, the proposed method lays a

foundation for next-generation ASV security, bridging effective raw-audio modeling with an adaptable design to counter evolving logical access threats.

Data availibility

We have used the publicly available datasets, i.e., ASVspoof2019 and ASVspoof 2021, which can be found at [31] and [32] in the paper.

Received: 18 March 2025; Accepted: 29 May 2025
Published online: 01 July 2025

References

1. Khan, A., Malik, K. M., Ryan, J. & Saravanan, M. Battling voice spoofing: A review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures. *J. Inf. Secur.* **45**, 123–156 (2023).
2. Wu, H., Liu, A. T. & Lee, H. Y. Defense for black-box attacks on anti-spoofing models by self-supervised learning. *IEEE/ACM Trans. Audio, Speech, and Language Process.* **28**, 1771–1783 (2020).
3. Delgado, H. *et al.* Asvspoof 2021: Automatic speaker verification spoofing and counter measures challenge evaluation plan. Preprint at arXiv:2109.00535 (2021).
4. Wu, Z. et al. Spoofing and counter measures for speaker verification: A survey. *Speech Commun.* **66**, 130–153 (2015).
5. D Alegre, F., Amehraye, A., Evans, N. Spoofing counter measures to protect automatic speaker verification from voice conversion. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 3068–3072 (IEEE, 2013).
6. Jati, A. et al. Adversarial attack and defense strategies for deep speaker recognition systems. *Comput. Speech & Language* **68**, 101199. https://doi.org/10.1016/j.csl.2021.101199 (2021).
7. Chakroborty, S. & Saha, G. Improved text-independent speaker identification using fused mfcc & imfcc feature sets based on gaussian filter. *Int. J. Signal Process.* **5**, 11–19 (2009).
8. Todisco, M., Delgado, H. & Evans, N. W. D. Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Comput. Speech & Language* **45**, 516–535 (2017).
9. Sahidullah, M. et al. *Integrated spoofing counter measures and automatic speaker verification: An evaluation on asvspoof 2015* (In Handbook of Biometric Anti-Spoofing (Springer, New York, 2016).
10. Korshunov, P. *et al.* Overview of btas 2016 speaker anti-spoofing competition. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 1–6 (IEEE, 2016).
11. Nagarsheth, P., el Khoury, E., Patil, K., Garland, M. Replay attack detection using dnn for channel discrimination. In *INTERSPEECH* (2017).
12. Lai, C.I. *et al.* Attentive filtering networks for audio replay attack detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6316–6320 (IEEE, 2019).
13. Wu, X., He, R., Sun, Z. & Tan, T. A light cnn for deep face representation with noisy labels. *IEEE Trans. Inf. Forensics Secur.* **13**, 2884–2896 (2018).
14. Tak, H. *et al.* End-to-end anti-spoofing with rawnet2. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6369–6373 (2021).
15. Ma, Y., Ren, Z., Xu, S. Rw-resnet: A novel speech anti-spoofing model using raw waveform. Preprint at arXiv:2108.05684https://doi.org/10.48550/ARXIV.2108.05684 (2021).
16. Li, X. *et al.* Replay and synthetic speech detection with res2net architecture. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6354–6358 (IEEE, 2021).
17. Arif, T., Javed, A., Alhameed, M., Jeribi, F. & Tahir, A. Voice spoofing countermeasure for logical access attacks detection. *IEEE Access* **9**, 162857–162868. https://doi.org/10.1109/ACCESS.2021.3133134 (2021).
18. Wang, X. et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Comput. Speech & Language* **64**, 101114 (2020).
19. Saito, Y., Takamichi, S. & Saruwatari, H. Statistical parametric speech synthesis incorporating generative adversarial networks. *IEEE/ACM Trans. Audio, Speech, and Language Process.* **26**, 84–96. https://doi.org/10.1109/TASLP.2017.2761547 (2018).
20. Chen, X., Zhang, Y., Zhu, G., Duan, Z. Ur channel-robust synthetic speech detection system for asvspoof 2021. In *Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Counter Measures Challenge*, 75–82, https://doi.org/10.21437/ASVSPOOF.2021-12 (2021).
21. Kinnunen, T. *et al.* The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In *INTERSPEECH*, https://doi.org/10.21437/Interspeech.2017-1111 (2017).
22. Xue, J. *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)*, 1–5, https://doi.org/10.1109/ICASSP49357.2023.10096837 (2023).
23. Chen, F., Deng, S., Zheng, T., He, Y. & Han, J. Graph-based spectro-temporal dependency modeling for anti-spoofing. In *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5, https://doi.org/10.1109/ICASSP49357.2023.10096741 (2023).
24. Ding, S., Zhang, Y. & Duan, Z. Samo: Speaker attractor multi-center one-class learning for voice antispoofing. In *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5, https://doi.org/10.1109/ICASSP49357.2023.10094704 (2023).
25. Javed, A., Malik, K. M., Irtaza, A. & Malik, H. Towards protecting cyber-physical and iot systems from single-and multi-order voice spoofing attacks. *Appl. Acoust.* **183**, 108283 (2021).
26. Jung, J.W. *et al.* Sasv 2022: The first spoofing-aware speaker verification challenge. Preprint at arXiv:2203.14732 (2022).
27. Zhang, Y., Zhu, G., Duan, Z. A probabilistic fusion framework for spoofing aware speaker verification. Preprint at arXiv:2202.05253 (2022).
28. Liu, X., Sahidullah, M., Kinnunen, T. Spoofing-aware speaker verification with unsupervised domain adaptation. Preprint at arXiv:2203.10992 (2022).
29. Todisco, M. *et al.* Asvspoof 2019: Future horizons in spoofed and fake audio detection. Preprint at arXiv:1904.05441 (2019).
30. Baumann, R. et al. Voice spoofing detection corpus for single and multi-order audio replays. *Comput Speech & Language* **65**, 101132 (2021).
31. Kinnunen, T. *et al.* T-dcf: A detection cost function for the tandem assessment of spoofing counter measures and automatic speaker verification. Preprint at arXiv:1804.09618 (2018).
32. Todisco, M. *et al.* Asvspoof 2019 logical access (la) dataset. https://datashare.ed.ac.uk/handle/10283/3336 (2019). Accessed on March 18, 2025.
33. Yamagishi, J. *et al.* Asvspoof 2021 logical access (la) dataset. https://zenodo.org/records/4837263 (2021). Accessed on March 18, 2025.
34. Yamagishi, J. *et al.* Asvspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *ASV Spoof***13** (2019).

35. Kumar, S. R. & Bharathi, B. A novel approach towards generalization of countermeasure for spoofing attack on asv systems. *Circuits Syst. Signal Process.* **40**, 872–889 (2021).
36. Monteiro, J., Alam, J. & Falk, T. H. Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers. *Comput. Speech & Language* **63**, 101096 (2020).
37. Lavrentyeva, G. *et al.* Audio replay attack detection with deep learning frameworks. In *Proceedings of Interspeech*, 82–86 (2017).
38. Javed, A., Malik, K. M., Malik, H. & Irtaza, A. Voice spoofing detector: A unified anti-spoofing framework. *Expert Syst. Appl.* **198**, 116770 (2022).
39. Tak, H. *et al.* End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. Preprint at arXiv:2107.12710https://doi.org/10.48550/ARXIV.2107.12710 (2021).
40. Arif, T., Javed, A., Alhameed, M., Jeribi, F. & Tahir, A. Voice spoofing countermeasure for logical access attacks detection. *IEEE Access* **9**, 162857–162868. https://doi.org/10.1109/ACCESS.2021.3133134 (2021).
41. Ravanelli, M., Bengio, Y. Speaker recognition from raw waveform with sincnet. In *IEEE Spoken Language Technology Workshop (SLT)* (2018).
42. Hendrycks, D., Gimpel, K. Gaussian error linear units (gelus). Preprint at arXiv:1606.08415 (2016).
43. Vaswani, A. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008 (2017).

Acknowledgements

The authors extend their appreciation to the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) for funding this work through (grant number IMSIU-DDRSP2504).

Author contributions

Conceptualization, H.A. and A.J.; methodology, H.A., A.J., T.K; software, H.A.; validation, A.J., A.K.J.S. and H.A.; formal analysis, H.A., A.J.; investigation, H.A., A.J., A.K.J.S. and T.K; resources, H.A., A.J.; data curation, H.A.; writing—original draft preparation, H.A., A.J.; writing—review and editing, H.A., A.J., T.K, and A.K.J.S.; visualization, H.A. and A.K.J.S.; supervision, A.J; project administration, A.J. and A.K.J.S.; funding acquisition, A.K.J.S.

Funding

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2504).

Declarations

Competing interests

The authors declare no conflicts of interest

Additional information

Correspondence and requests for materials should be addressed to A.K.J.S.