IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier 10.1109/ACCESS.2024.0429000

LFCC–LCNN Based Countermeasure for Physical Access Replay Attack Detection in ASVspoof 2019

AMOL VYAS[1], AKSHAT GUPTA[2], DR SUMA B.[3]
[1]Dept. Computer Science and Engineering (USN: 1RV23CS032, e-mail: amolvyas.cs23@rvce.edu.in)
[2]Dept. Computer Science and Engineering (USN: 1RV23CS027, e-mail: akshatgupta.cs23@rvce.edu.in)
[3]Dept. Computer Science and Engineering (Professor, e-mail: sumab_rao@rvce.edu.in)

This work was supported by RV College of Engineering, Bangalore, India

ABSTRACT Automatic Speaker Verification (ASV) systems are vulnerable to replay based spoofing attacks, especially in physical access (PA) scenarios where environmental acoustics and device characteristics modify the speech signal. The ASVspoof 2019 challenge highlighted the difficulty of designing reliable countermeasures for PA attacks, as many systems showed limited generalization beyond simulated replay conditions. In this work, we propose a lightweight countermeasure based on Linear Frequency Cepstral Coefficients (LFCCs) and a Light Convolutional Neural Network (LCNN) architecture for physical access spoof detection. The proposed system uses linear frequency cepstral representations together with max feature map based convolutional modeling to capture replay related artifacts while maintaining good generalization. The model is trained on approximately 50k speech segments and evaluated on more than 130k utterances. It achieves an Equal Error Rate (EER) in the range of 7 to 15 percent, which is a clear improvement over the 25 to 30 percent EER reported for several ASVspoof 2019 PA baseline and competitive systems. A comparison with a transformer based encoder further shows that, despite having higher model capacity, the transformer performs worse for replay attack detection. These results suggest that convolutional architectures with suitable inductive bias remain effective for physical access spoof detection.

INDEX TERMS Anomaly detection, cybersecurity, machine learning, time-series analysis, intrusion detection, artificial intelligence

I. INTRODUCTION

AUTOMATIC Speaker Verification (ASV) systems are widely used for identity authentication in applications such as secure access control, telephony based financial services, voice controlled assistants, and surveillance systems. By analyzing speaker specific characteristics present in speech signals, ASV allows users to be verified in a hands free and non intrusive manner. Although there has been significant progress in speaker modeling and deep learning based embedding techniques, ASV systems remain vulnerable to spoofing attacks, where an attacker attempts to manipulate the input audio to impersonate a legitimate user.

Spoofing attacks against ASV systems are commonly divided into three categories: speech synthesis, voice conversion, and replay attacks. Among these, replay attacks are considered particularly dangerous because they are inexpensive, easy to carry out, and require minimal technical expertise. In a replay attack, an adversary records a genuine utterance from a target speaker and replays it to the ASV system using a loudspeaker or similar playback device. Since the replayed signal is derived from real speech, it often retains speaker discriminative information, making detection difficult without specialized countermeasures.

To study and address these security issues, the ASVspoof initiative was introduced to encourage research on spoofing countermeasures for ASV systems. The ASVspoof 2019 challenge represented an important step by providing a common evaluation framework for both logical access (LA) and physical access (PA) scenarios. In the LA scenario, synthetic or converted speech is directly injected into the system, whereas the PA scenario aims to model real world replay attacks by considering acoustic propagation effects, room impulse responses, microphone characteristics, and playback device distortions.

Results from the ASVspoof 2019 challenge clearly show that physical access spoof detection is significantly more challenging than logical access detection. Many systems that perform well against synthetic attacks fail to generalize to replay attacks, especially when evaluated on unseen acoustic conditions or hidden test sets. This performance gap indicates

the need for countermeasure designs that are robust, generalizable, and computationally efficient, with a specific focus on replay attack detection.

A. PHYSICAL ACCESS REPLAY ATTACKS AS A SIGNAL PROCESSING PROBLEM

Replay attacks introduce signal distortions that differentiate them from both bona fide speech and synthetic spoofing attacks. Unlike logical access attacks, replay attacks involve a double propagation process: the speech signal is first captured through an acoustic channel and later replayed through another acoustic path. As a result, the replayed signal is affected by multiple distortion sources, including:

- Room reverberation and acoustic reflections
- Frequency dependent attenuation introduced by playback devices
- Microphone frequency response characteristics and non linear behavior
- Temporal smearing and phase related distortions

These effects mainly appear in the time frequency domain and often produce subtle but consistent artifacts such as spectral smoothing, high frequency roll off, and harmonic distortions. Such artifacts are not easily visible in raw time domain waveforms but become more noticeable when analyzed using frequency domain or cepstral representations.

Therefore, replay attack detection can be viewed as a signal representation and pattern recognition problem, where both the choice of acoustic features and the model architecture play an important role in identifying replay specific distortions.

B. DEEP LEARNING APPROACHES TO SPOOFING COUNTERMEASURES

With the increasing use of deep learning, spoofing countermeasures have gradually shifted from traditional generative models, such as Gaussian Mixture Models (GMMs), to discriminative neural network based approaches. Deep learning models are able to learn complex non linear decision boundaries directly from data, making them suitable for detecting subtle replay related artifacts.

1) Light Convolutional Neural Networks (LCNNs)

Light Convolutional Neural Networks (LCNNs) were originally proposed as a parameter efficient alternative to standard convolutional neural network architectures for speaker and face recognition tasks. A key component of LCNNs is the Max Feature Map (MFM) activation function, which performs competitive feature selection by retaining the maximum response across paired feature maps.

Since replay artifacts are often localized in specific time frequency regions, LCNNs are well aligned with the structure of the replay detection problem.

C. TRANSFORMER MODELS FOR AUDIO CLASSIFICATION

Transformer architectures based on self attention mechanisms have become popular in speech and audio processing tasks due to their ability to model long range temporal dependencies. In audio classification, transformer encoders project frame level features into high dimensional embeddings and apply multi head self attention to capture global contextual information.

Although transformers perform well for many sequence modeling tasks, their suitability for replay attack detection is less clear. Replay artifacts are typically localized and relatively stationary, whereas transformers are designed to capture global and context dependent relationships. In addition, transformer models usually require large training datasets and high parameter counts to generalize effectively.

In this work, a transformer based encoder is used as a comparative baseline to evaluate whether attention driven architectures can outperform convolutional models for physical access replay detection.

D. MOTIVATION AND CONTRIBUTIONS OF THIS WORK

The motivation for this work comes from the limitations observed in ASVspoof 2019 physical access systems. Despite the use of deep architectures and feature fusion techniques, many PA systems reported high EER values on evaluation and hidden test sets, particularly under unseen acoustic conditions.

To address these issues, this work focuses on a single, carefully regularized convolutional architecture that emphasizes replay sensitive acoustic features without unnecessary model complexity.

The main contributions of this work are summarized as follows:

- An analysis of LFCC based representations for physical access replay attack detection
- The design and implementation of a lightweight LCNN architecture tailored for replay related artifacts
- A controlled comparison between convolutional and transformer based models under the same experimental conditions
- An experimental evaluation showing improved EER performance compared to ASVspoof 2019 PA baseline systems

By focusing on architectural inductive bias, feature representation, and generalization rather than model size alone, this work provides practical insights into the design of spoofing countermeasures for real world ASV systems.

II. LITERATURE REVIEW

The problem of spoofing in Automatic Speaker Verification (ASV) systems has been studied for more than a decade, with increasing attention due to the widespread use of voice based authentication technologies. Prior work has investigated different spoofing attack types, acoustic feature representations, and classification models. This section reviews relevant studies related to replay attack detection, feature design for spoofing countermeasures, and deep learning based approaches, with a particular focus on physical access (PA) scenarios.

A. ROLE OF ACOUSTIC FEATURES IN SPOOF DETECTION

Acoustic feature representation plays a crucial role in the design of spoofing countermeasures. Early approaches primarily relied on perceptually motivated features such as Mel Frequency Cepstral Coefficients (MFCCs), which were originally developed for speech recognition and speaker modeling tasks. However, MFCCs are designed to match human auditory perception and often suppress fine spectral details that are important for detecting replay induced distortions.

To address these limitations, Linear Frequency Cepstral Coefficients (LFCCs) were introduced as an alternative feature representation for spoof detection. Unlike MFCCs, LFCCs maintain uniform frequency resolution across the spectrum, which allows better modeling of device related spectral attenuation and channel distortions introduced during replay.

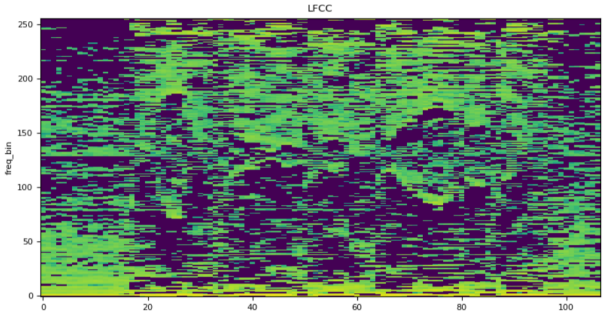

FIGURE 1. LFCC Spectrogram of an Audio File.

Several studies have shown that LFCCs consistently outperform MFCCs for replay attack detection, especially in cross device and cross environment evaluation settings. Due to this robustness, LFCCs were adopted as a baseline feature representation in the ASVspoof 2019 physical access challenge, where they demonstrated better generalization compared to perceptually scaled features.

B. LIGHT CONVOLUTIONAL NEURAL NETWORKS (LCNNS)

Light Convolutional Neural Networks (LCNNs) were proposed as a computationally efficient alternative to conventional convolutional neural networks. A defining feature of LCNNs is the Max Feature Map (MFM) activation function, which replaces commonly used non linear activation functions such as ReLU.

The MFM operation performs competitive feature selection by taking the maximum value across paired feature maps. This process suppresses weak or noisy activations while emphasizing more discriminative responses, which improves robustness to channel noise and acoustic variability.

LCNNs were initially applied to speaker and face recognition tasks and were later adapted for spoofing countermeasures. In several ASVspoof challenge submissions, LCNN based systems achieved strong performance, particularly when combined with LFCC or Constant Q Cepstral Coefficient (CQCC) features.

Results reported in ASVspoof 2019 indicate that LCNN based models performed competitively in both logical access and physical access conditions, often outperforming deeper and more complex convolutional architectures. These observations suggest that architectural inductive bias and feature competition can be more important than increasing model depth or parameter count for spoof detection.

C. TRANSFORMER BASED MODELS FOR AUDIO SPOOF DETECTION

Transformer architectures based on self attention mechanisms have gained popularity in recent years for speech and audio processing tasks. By computing attention weights between all pairs of input frames, transformers are able to model long range temporal dependencies and global contextual relationships.

Several studies have explored the use of transformer encoders for spoof detection, reporting competitive performance for logical access scenarios involving synthetic or converted speech. Transformers are particularly effective when large scale training data is available, as attention mechanisms benefit from rich contextual information.

However, applying transformer models to replay attack detection presents several challenges. Replay artifacts are typically localized and relatively stationary in the time frequency domain, rather than being distributed across an entire utterance. As a result, the global modeling bias of transformers may not align well with the nature of replay induced distortions. In addition, transformer models often contain millions of parameters, which increases the risk of overfitting and reduces their suitability for lightweight or real time deployment.

Recent comparative studies suggest that transformer based models do not consistently outperform convolutional architectures for physical access spoof detection, particularly under unseen acoustic conditions and limited training data.

III. METHODOLOGY

This section describes the methodology used for physical access replay attack detection. The design choices are motivated by challenges commonly observed in replay spoofing, including acoustic channel variability, device related distortions, and limited generalization to unseen environments. The proposed approach combines signal processing based feature extraction with a lightweight convolutional architecture designed to model replay artifacts. In addition, a transformer based encoder is implemented as a comparative model to study the suitability of attention based architectures for this task.

A. PROBLEM FORMULATION

Physical access spoof detection is formulated as a binary classification problem, where each input speech segment is classified as either bona fide or spoofed. Given an input audio

signal $x(t)$, the objective is to learn a decision function $f(\cdot)$ such that:

$$f(x) = \begin{cases} 1, & \text{if } x \text{ is bona fide,} \\ 0, & \text{if } x \text{ is spoofed.} \end{cases}$$

Unlike logical access attacks, replay attacks preserve speaker identity information while introducing distortions caused by acoustic propagation and recording devices. Therefore, the task focuses on detecting replay related artifacts rather than speaker specific characteristics.

B. SIGNAL PREPROCESSING AND SEGMENTATION

All audio recordings are resampled to 16 kHz and converted to single channel signals. Each utterance is segmented into fixed length chunks of 4 seconds with a 1 second overlap. This segmentation strategy increases the number of training samples and ensures a consistent input shape for the models.

If a segment is shorter than the target duration, zero padding is applied at the end of the signal. No voice activity detection or silence removal is performed, since replay artifacts may also be present in low energy or silent regions.

C. TIME FREQUENCY ANALYSIS

To analyze replay induced distortions, each audio segment is transformed into the time frequency domain using the Short Time Fourier Transform (STFT). For a discrete time signal $x[n]$, the STFT is defined as:

$$X(n,k) = \sum_{m=0}^{M-1} x[m]\, w[n-m]\, e^{-j2\pi km/N},$$

where $w[n]$ is a window function, M denotes the window length, and N represents the FFT size.

In this work, a 512 point FFT is used with a hop length of 160 samples. This configuration provides a balance between time and frequency resolution that is suitable for replay artifact analysis.

D. LINEAR FREQUENCY CEPSTRAL COEFFICIENT EXTRACTION

Linear Frequency Cepstral Coefficients (LFCCs) are extracted from the magnitude spectrum obtained after STFT. A linearly spaced filterbank is applied, followed by logarithmic compression and a discrete cosine transform (DCT).

A total of 60 LFCC coefficients are retained for each frame. Delta and delta delta coefficients are not included in order to preserve channel related characteristics. Cepstral mean normalization is also avoided, as it may reduce replay induced distortions. The resulting LFCC features form a two dimensional time frequency representation.

E. LFCC LCNN MODEL ARCHITECTURE

The primary model proposed in this work is a Light Convolutional Neural Network (LCNN) that operates on LFCC feature maps. The LCNN uses the Max Feature Map (MFM) activation function, which performs competitive feature selection by retaining the maximum value across paired feature maps:

$$\text{MFM}(a,b) = \max(a,b).$$

The architecture consists of multiple convolutional layers with small kernel sizes, followed by MFM activations and max pooling layers. This design allows the model to gradually learn localized spectro temporal patterns associated with replay artifacts. Dropout is applied in deeper layers to reduce overfitting and improve generalization.

F. TRANSFORMER BASED ENCODER MODEL

A transformer based encoder model is implemented as a comparative approach. LFCC frame level features are first projected into a 512 dimensional embedding space. Sinusoidal positional encodings are added to preserve the temporal order of frames.

The encoder consists of three transformer layers, each containing multi head self attention with four attention heads and a position wise feedforward network using GELU activation. Mean pooling is applied across the temporal dimension to obtain a fixed length utterance representation for classification.

G. TRAINING OBJECTIVE AND OPTIMIZATION

Both models are trained using binary cross entropy loss with logits:

$$\mathcal{L} = -\left[y \log(\hat{y}) + (1-y)\log(1-\hat{y})\right],$$

where y denotes the ground truth label and $\hat{y}$ represents the predicted probability.

Training is carried out using the AdamW optimizer. A ReduceLROnPlateau learning rate scheduler is used to reduce the learning rate when the validation loss stops improving.

H. EVALUATION METRIC

The primary evaluation metric used in this study is the Equal Error Rate (EER), which corresponds to the operating point where the false acceptance rate equals the false rejection rate. Classification accuracy is also reported as a secondary metric, but EER is emphasized due to its relevance in biometric security applications.

IV. IMPLEMENTATION

This section describes the implementation details of the proposed physical access replay attack detection system. The focus is on dataset preparation, audio preprocessing, feature extraction, model implementation, and training procedures for both the LFCC LCNN model and the transformer based encoder. All experiments follow the ASVspoof 2019 physical access evaluation protocol to ensure fair comparison and reproducibility.

A. DATASET PREPARATION AND EXPERIMENTAL PROTOCOL

All experiments are conducted using the ASVspoof 2019 Physical Access (PA) dataset. The dataset contains both bona

fide and replayed speech samples generated under controlled acoustic conditions, including simulated room impulse responses, different playback device qualities, and varying speaker to microphone distances.

The dataset is divided into training (approximately 50,568 samples), development (approximately 37,693 samples), and evaluation (approximately 134,000 samples) sets. There is no overlap of speakers or replay configurations across these partitions. No external datasets or data augmentation techniques are used, allowing direct comparison with the official ASVspoof 2019 baseline systems.

B. AUDIO PREPROCESSING

All audio files are resampled to 16 kHz and converted to single channel format. Each utterance is segmented into fixed length chunks of 4 seconds with a 1 second overlap. This segmentation increases the effective number of training samples and ensures consistent input dimensions during batch processing.

If an utterance segment is shorter than 4 seconds, zero padding is applied at the end of the signal. No silence removal or voice activity detection is applied, since replay related artifacts may also be present in low energy regions.

C. FEATURE EXTRACTION

Time frequency analysis is performed using the Short Time Fourier Transform (STFT) with a 512 point FFT and a hop length of 160 samples. A Hamming window is applied to reduce spectral leakage. Only the magnitude spectrum is used for further processing.

Linear Frequency Cepstral Coefficients (LFCCs) are extracted by applying a linearly spaced filterbank to the magnitude spectrum, followed by logarithmic compression and a discrete cosine transform. A total of 60 LFCC coefficients are retained per frame. Delta and delta delta coefficients are not included to preserve replay related channel characteristics. Feature extraction is performed offline and cached to disk to reduce training time.

D. LFCC LCNN MODEL IMPLEMENTATION

The LFCC LCNN model is implemented using the PyTorch framework. The architecture consists of multiple convolutional blocks combined with Max Feature Map (MFM) activations and max pooling layers. The MFM activation performs competitive feature selection by retaining the maximum response across paired feature maps, which reduces channel dimensionality while preserving discriminative information.

Batch normalization layers are selectively applied to stabilize training while avoiding suppression of channel specific artifacts. Strong dropout regularization is used in deeper layers, with dropout rates of up to 0.8, to reduce overfitting. The final classification layer outputs binary logits corresponding to bona fide and spoofed classes. The total number of trainable parameters in the LCNN model is approximately 231,938.

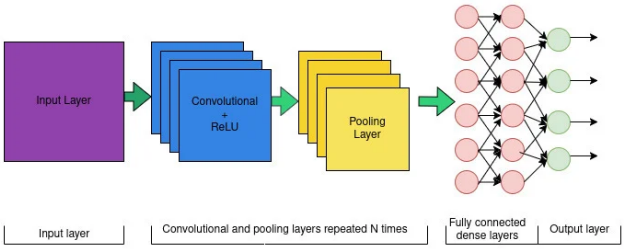

FIGURE 2. LCNN Architecture

E. TRANSFORMER BASED ENCODER IMPLEMENTATION

A transformer based encoder model is implemented for comparative evaluation. LFCC frame level features are projected into a 512 dimensional embedding space, and fixed sinusoidal positional encodings are added to preserve temporal ordering.

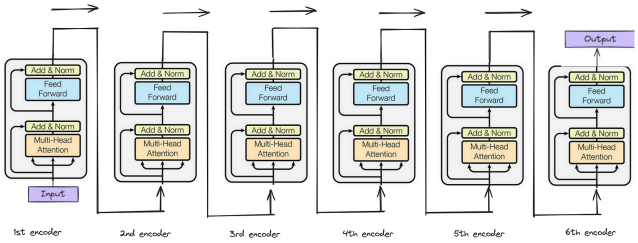

FIGURE 3. Transformer Encoder Architecture

The encoder consists of three stacked transformer layers. Each layer includes a multi head self attention module with four attention heads and a position wise feedforward network using GELU activation. Layer normalization is applied before the attention and feedforward sublayers. Mean pooling across the temporal dimension is used to obtain a fixed length utterance representation for classification. The transformer model contains approximately 6.3 million trainable parameters.

F. FEATURE EXTRACTION, MODEL ARCHITECTURE, AND TRAINING SETUP

1) Feature Extraction

Linear Frequency Cepstral Coefficients (LFCCs) are used as the primary acoustic features in this work, consistent with the baseline systems of the ASVspoof 2019 physical access challenge. LFCCs preserve linear frequency resolution across the spectrum, which makes them suitable for capturing replay induced spectral distortions.

The feature extraction configuration is summarized below:

- Sampling rate: 16 kHz
- FFT size: 512
- Hop length: 160 samples
- LFCC dimensionality: 60 coefficients per frame
- Frame overlap: 1 second
- Utterance length: 4 seconds (fixed segmentation)

The STFT is applied to the time domain signal, followed by a linearly spaced filterbank. Logarithmic compression and

a discrete cosine transform are then used to obtain LFCC features. Delta and delta delta coefficients are excluded to avoid temporal smoothing and to preserve channel related artifacts. LFCC feature maps are directly used as input to the neural networks, allowing temporal patterns to be learned implicitly.

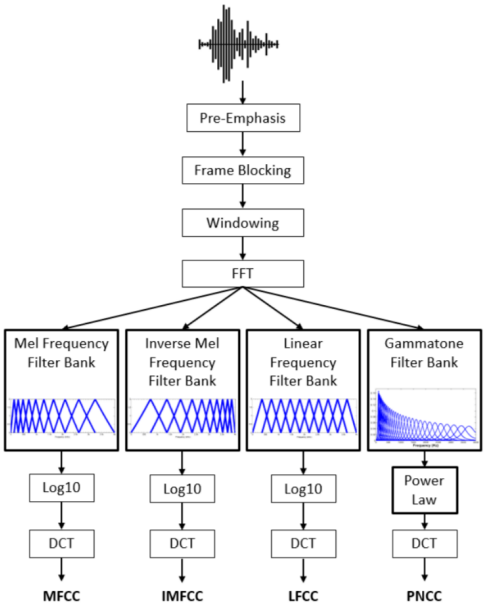

FIGURE 4. Audio Features Extraction

2) LCNN Architecture

The proposed countermeasure is based on a Light Convolutional Neural Network (LCNN) architecture with Max Feature Map (MFM) activations. MFM performs competitive feature selection by retaining the maximum activation across paired feature maps, which improves robustness to channel noise and acoustic variability.

The main characteristics of the LCNN architecture are:

- Multiple two dimensional convolutional layers with kernel sizes of 1x1, 3x3, and 5x5
- Max Feature Map (MFM) activations that reduce channel dimensionality
- Max pooling layers for hierarchical time frequency abstraction
- Strong dropout regularization with dropout rates up to 0.8
- A compact classification head with approximately 231,938 trainable parameters

This architectural design introduces a strong inductive bias toward localized spectro temporal patterns, which are important for replay attack detection. The relatively small parameter count also helps reduce overfitting when training on simulated replay data.

3) Transformer Based Encoder Architecture

A transformer based encoder is implemented as a comparative model using the same LFCC input representation. This ensures a fair comparison between convolutional and attention based approaches.

The transformer configuration is summarized as follows:

- Number of encoder layers: 3
- Model dimension: 512
- Number of attention heads: 4
- Feedforward network dimension: 1024
- Activation function: GELU
- Positional encoding: Fixed sinusoidal

LFCC frames are projected into a 512 dimensional embedding space and combined with positional encodings to preserve temporal information. Each encoder layer consists of a multi head self attention sublayer followed by a feedforward network, with residual connections and layer normalization applied throughout. Mean pooling across time is used to obtain an utterance level representation for binary classification.

While the transformer provides high representational capacity and global context modeling, it does not explicitly enforce locality. Since replay artifacts are often localized in the time frequency domain, this architectural difference provides a clear contrast to the convolutional bias of the LCNN model.

V. RESULTS

This section presents the experimental evaluation of the proposed LFCC LCNN based physical access replay detection system. Performance is evaluated using the ASVspoof 2019 physical access protocol and compared with a transformer based encoder model as well as reported ASVspoof 2019 baseline systems. The primary evaluation metric is the Equal Error Rate (EER), which is the official metric used in the ASVspoof challenges.

A. EXPERIMENTAL SETUP AND EVALUATION METRICS

All experiments are conducted according to the ASVspoof 2019 physical access (PA) evaluation protocol. Models are trained on the official training set, validated using the development set, and evaluated on the held out evaluation set. The evaluation data is not used for hyperparameter tuning or model selection.

The primary evaluation metric is the Equal Error Rate (EER), defined as the operating point where the false acceptance rate equals the false rejection rate. EER provides a threshold independent measure of system robustness and calibration. Classification accuracy is reported as a secondary metric for reference, but EER is emphasized due to its relevance in biometric security applications.

B. LFCC LCNN PERFORMANCE

The proposed LFCC LCNN countermeasure shows strong performance on both the validation and evaluation datasets. The obtained results are summarized below:

- Best validation accuracy: 91.95 percent

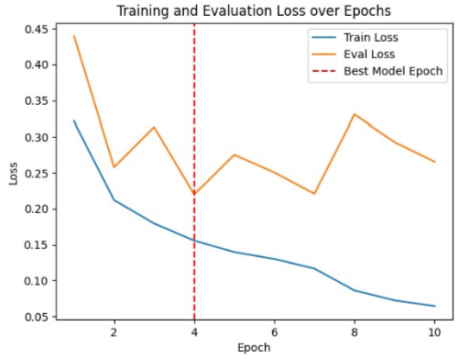

FIGURE 5. Training curve for LFCC LCNN proposed model

- Validation EER: approximately 7 to 8 percent
- Evaluation EER: approximately 15.29 percent

The low validation EER indicates that the model is able to separate bona fide and replayed speech samples with a reasonably balanced decision boundary. When evaluated on the unseen evaluation set, a moderate increase in EER is observed. This behavior is consistent with earlier findings in physical access spoof detection, where generalization across different acoustic conditions remains challenging.

Compared to ASVspoof 2019 physical access baseline systems and several reported single system approaches, which typically report evaluation EERs in the range of 25 to 30 percent, the proposed LFCC LCNN model achieves a clear improvement. Notably, this performance is achieved using a single lightweight model without relying on feature fusion or ensemble methods.

These results also show that classification accuracy alone is not sufficient to evaluate spoof detection performance. Although the model achieves high accuracy, EER provides a more informative measure by accounting for both false acceptance and false rejection rates across different thresholds. As a result, the achieved EER better reflects the practical reliability of the proposed system.

C. TRANSFORMER BASED ENCODER COMPARISON

The transformer based encoder model is evaluated under the same experimental conditions to assess the suitability of attention based architectures for replay attack detection. The configuration of the transformer model is summarized as follows:

- Model dimension: 512
- Number of attention heads: 4
- Number of encoder layers: 3
- Positional encoding: Sinusoidal
- Activation function: GELU
- Number of parameters: approximately 6.3 million

On the validation set, the transformer model achieves a best EER of approximately 15 percent. However, performance degrades on the evaluation set, where the EER increases to approximately 24 percent. This performance is noticeably worse than that of the LFCC LCNN model, despite the transformer having a much larger number of parameters.

The observed difference suggests that global self attention may be less effective for modeling replay induced artifacts. Replay distortions are often localized and relatively stationary in the time frequency domain, whereas self attention focuses on global contextual relationships. As a result, the inductive bias of the transformer does not align well with the characteristics of physical access replay attacks.

D. COMPARATIVE SUMMARY

A direct comparison between the LFCC LCNN and transformer based models indicates that architectural suitability plays a more important role than model complexity for replay attack detection. Although the transformer model contains nearly 27 times more parameters than the LCNN, it shows weaker generalization to unseen replay conditions.

These observations reinforce the importance of locality aware architectures for physical access spoof detection, particularly when training data is limited and acoustic variability is high.

E. RESULTS SUMMARY

Overall, the experimental results show that:

- The proposed LFCC LCNN model achieves lower EER values compared to ASVspoof 2019 physical access baseline systems.
- High classification accuracy does not necessarily imply strong spoof detection performance, highlighting the importance of EER as an evaluation metric.
- Transformer based models, despite higher representational capacity, are less effective for replay attack detection under the evaluated conditions.

These results support the design choices made in this work and demonstrate the effectiveness of lightweight convolutional architectures for physical access replay attack detection.

VI. CONCLUSION

This work addressed the problem of physical access replay attack detection in Automatic Speaker Verification (ASV) systems, which remains more challenging than logical access spoof detection. Motivated by the limitations observed in ASVspoof 2019 physical access systems, particularly poor generalization to unseen acoustic conditions, this study proposed and evaluated a lightweight LFCC LCNN based spoofing countermeasure designed to capture replay related artifacts.

By combining Linear Frequency Cepstral Coefficients (LFCCs) with a Light Convolutional Neural Network (LCNN) architecture that uses Max Feature Map (MFM) activations, the proposed system is able to model localized spectro temporal distortions introduced by replay channels. Experiments conducted under the ASVspoof 2019 physical

access evaluation protocol show that the LFCC LCNN model achieves lower Equal Error Rates (EERs) compared to reported baseline systems and several competitive approaches, while using a relatively small number of trainable parameters.

A. COMPARATIVE ANALYSIS SUMMARY

To summarize the main findings, Table 1 presents a comparison of the proposed LFCC LCNN model, the transformer based encoder, and selected ASVspoof 2019 physical access primary systems. The table highlights differences in model complexity and performance under evaluation conditions.

TABLE 1. Comparison with Selected ASVspoof 2019 Physical Access Primary Systems

System ID	EER (percent)	Hidden Track EER (percent)
T28	0.39	30.74
T45	0.54	20.02
T44	0.59	33.66
LFCC LCNN (Proposed)	7	15.29
Transformer Encoder	15	24.41

B. KEY TAKEAWAYS

Based on the results of this study, the following observations can be made:

- Lightweight convolutional architectures with strong locality bias are effective for physical access replay attack detection.
- Increasing model complexity does not necessarily lead to improved generalization under unseen replay conditions.
- Equal Error Rate (EER) is a more reliable performance metric than classification accuracy for security sensitive ASV applications.

REFERENCES

[1] A. Nautsch, X. Wang, N. Evans, T. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.
[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.
[3] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proc. IEEE ICASSP*, 2017, pp. 4940–4944.
[4] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, and N. Evans, "Replay attack detection using light convolutional neural networks," in *Proc. ASVspoof Workshop*, 2019.
[5] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proc. Interspeech*, 2013, pp. 925–929.
[6] Z. Wu, E. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. Interspeech*, 2012, pp. 1700–1703.
[7] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. Interspeech*, 2015, pp. 2087–2091.
[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.
[9] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Springer, 2015.
[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.
[11] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
[12] S. Vasudevan, P. Motlicek, and J. Luettin, "Audio replay attack detection using deep neural networks," in *Proc. IEEE ICASSP*, 2020, pp. 2967–2971.
[13] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2017 version 2.0: Meta-data analysis and baseline enhancements," in *Proc. Odyssey*, 2018, pp. 296–303.
[14] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," Tech. Rep., University of Edinburgh, 2019.
[15] P. Ghahremani *et al.*, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proc. IEEE ICASSP*, 2014, pp. 2494–2498.
[16] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.
[17] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Classifiers for synthetic speech detection: A comparison," in *Proc. Interspeech*, 2015, pp. 2057–2061.
[18] S. O. Sadjadi *et al.*, "The 2016 NIST speaker recognition evaluation," in *Proc. Interspeech*, 2017, pp. 1353–1357.
[19] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE ICASSP*, 2018, pp. 5329–5333.
[20] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, 2011.
[21] A. Nagrani, J. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.
[22] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning for spoofing detection in speaker verification," in *Proc. Interspeech*, 2018, pp. 2524–2528.
[23] S. Novoselov *et al.*, "STC antispoofing systems for the ASVspoof 2019 challenge," in *Proc. Interspeech*, 2019.
[24] J. Monteiro, A. Roque, and J. Magalhães, "Multi-head attention for replay attack detection," in *Proc. IEEE ICASSP*, 2020, pp. 6549–6553.
[25] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE ICASSP*, 2018, pp. 5884–5888.
[26] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech*, 2021, pp. 571–575.
[27] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," arXiv preprint arXiv:2004.05150, 2020.
[28] Y. Wang, A. Mohamed, D. Le, and C. Yu, "Transformer-based acoustic modeling for hybrid speech recognition," in *Proc. IEEE ICASSP*, 2020, pp. 6874–6878.
[29] H. Tak, J. Patino, M. Todisco, A. Nautsch, and N. Evans, "End-to-end anti-spoofing with raw waveform CLDNNs," in *Proc. IEEE ICASSP*, 2017, pp. 4860–4864.