SpoTNet: A spoofing-aware Transformer Network for Effective Synthetic Speech Detection

Awais Khan
Khalid Mahmood Malik*
awaiskhan@oakland.edu
mahmood@oakland.edu
Department of Computer Science and Engineering, Oakland University, 532 EC 115 Library Drive, Rochester, MI 48309-4479, USA

ABSTRACT

The prevalence of voice spoofing attacks in today's digital world has become a critical security concern. Attackers employ various techniques, such as voice conversion (VC) and text-to-speech (TTS), to generate synthetic speech that imitates the victim's voice and gain access to sensitive information. The recent advances in synthetic speech generation pose a significant threat to modern security systems, while traditional voice authentication methods are incapable of detecting them effectively. To address this issue, a novel solution for logical access (LA)-based synthetic speech detection is proposed in this paper. SpoTNet is an attention-based spoofing transformer network that includes crafted front-end spoofing features and deep attentive features retrieved using the developed logical spoofing transformer encoder (LSTE). The derived attentive features were then processed by the proposed multi-layer spoofing classifier to classify speech samples as bona fide or synthetic. In synthetic speeches produced by the TTS algorithm, the spectral characteristics of the synthetic speech are altered to match the target speaker's formant frequencies, while in VC attacks, the temporal alignment of the speech segments is manipulated to preserve the target speaker's prosodic features. By highlighting these observations, this paper targets the prosodic and phonetic-based crafted features, i.e., the Mel-spectrogram, spectral contrast, and spectral envelope, presenting an effective preprocessing pipeline proven to be effective in synthetic speech detection. The proposed solution achieved state-of-the-art performance against eight recent feature fusion methods with lower EER of 0.95% on the ASVspoof-LA dataset, demonstrating its potential to advance the field of speaker identification and improve speaker recognition systems.

CCS CONCEPTS

• **Voice spoofing detection**; • **Synthetic Media Detection**; • **Multimedia forensics**;

KEYWORDS

Voice spoofing detection, Logical attacks, Speech synthesis, Text-to-speech, Voice conversion

ACM Reference Format:
Awais Khan and Khalid Mahmood Malik. 2023. SpoTNet: A spoofing-aware Transformer Network for Effective Synthetic Speech Detection. In *2nd ACM International Workshop on Multimedia AI against Disinformation (MAD '23), June 12–15, 2023, Thessaloniki, Greece.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3592572.3592841

1 INTRODUCTION

The advent of automation in voice technology has offered many advantages, but it has also raised serious security risks. The proliferation of voice spoofing attacks is one of the most significant challenges in voice-enabled devices such as Google Assistant, Siri, and Alexa. These attacks use various techniques to imitate the voice of legitimate speakers and gain unauthorized access to sensitive information. Voice spoofing attacks are divided into two categories: Physical Access (PA) and Logical Access (LA). Replay attacks, which use pre-recorded voice samples to generate synthetic speech, fall under PA, while LA uses text-to-speech (TTS) and voice conversion (VC) techniques to produce synthesized speech. Recent developments in TTS and VC techniques have enabled the generation of synthetic speech that is becoming increasingly indistinguishable from natural human speech. This, however, has led to an increase in LA attacks, which pose a significant threat to automatic speaker verification systems (ASVs). One of the most challenging issues facing ASVs is the inability to differentiate between genuine and synthetic speech samples. Recent research has highlighted the need for more advanced methods to detect voice spoofing attacks, particularly those that use TTS and VC techniques.

Detecting voice spoofing attacks relies on identifying irregularities in audio transmission caused by these attacks. These irregularities include artifacts such as ambient noise, compression artifacts, and microphone or speaker distortion for PA attacks, and phase mismatches, prosodic inconsistencies, and spectral disparities for LA attacks. Text-to-speech, voice conversion, and replay attacks can affect the spectral and temporal characteristics of speech signals, particularly in certain frequency or spectral regions [Bhangale et al. 2018; Huang et al. 2023]. For example, LA attacks induce spectral mismatches in higher frequency ranges, while PA attacks induce distortions across lower and higher time frames. To accurately detect these attacks, a comprehensive analysis of both spectral and temporal characteristics is necessary.

MAD '23, June 12–15, 2023, Thessaloniki, Greece

ACM ISBN 979-8-4007-0187-0/23/06…$15.00
https://doi.org/10.1145/3592572.3592841

Several approaches have been proposed to detect logical access attacks, including traditional feature-based methods, deep learning-based methods, and hybrid approaches. Recent studies have concentrated on developing deep learning-based solutions for audio spoofing detection that are more efficient and robust deep learning-based solutions for audio spoofing detection [Liu et al. 2022; Parmar et al. 2018] [Kulkarni and Khaparde 2017; Wolf et al. 2020]. While we claims that integrating effective handcrafted detection features and deep attentive features altogether might increase the accuracy and endurance of voice spoofing detection systems. In consequence, we propose SpoTNet, a spoofing-aware transformer network for effective synthetic speech detection in this paper. SpoTNet includes a novel logical spoofing transformer encoder (LSTE) that extracts attentive features from speech signals to differentiate between bona fide and synthetic speech samples. The proposed LSTE integrates traditional and deep learning-based methods to extract acoustic features that accurately capture the characteristics of bona fide and synthetic speech samples. The presented SpoTNet approach provides a responsive system for voice synthesis spoofing detection and increases the security of automatic speaker verification systems. Our approach can potentially serve as a template for further research in detecting voice spoofing attacks, particularly those using TTS and VC techniques. The main contribution of the paper are as follows:

- We present an effective acoustic feature set that effectively captures the traits of bona fide and synthetic speech samples by highlighting spectral and attentive artifacts in the audio signal.
- We present a novel spoofing transformer network (SpotNet) for detecting text-to-speech (TTS) and voice-conversion (VC) spoofing attacks using a proposed logical spoofing transformer encoder (LSTE).
- We provide rigorous experimentation on ASVspoof 2019 corpus to evaluate the significance of our proposed SpotNet spoofing detector over existing state-of-the-art feature fusion based recent techniques.
- The experimental results on ASVspoof2019 LA demonstrate that the proposed SpotNet approach outperforms existing state-of-the-art methodologies and provides a system responsive to voice synthesis spoofing detection, as well as the ability to increase the security of automatic speaker verification systems.

The remaining sections of the paper are structured as follows: In Section 2, the literature review is presented, in Section 3 the paper describes the proposed method, including the acoustic feature set, Logical Spoofing Transformer Encoder (LSTE) and spoofing multi layer classifier. Section 3 presents experimental results and comparisons with existing state-of-the-art techniques using the ASVspoof 2019 corpus. Finally, in Section 4, the paper concludes with a summary of the findings and contributions of the proposed method.

2 LITERATURE REVIEW

Several countermeasures have been proposed to counter voice spoofing attacks, and the traditional countermeasures are often comprised of two parts: the first one (front-end) is a feature representation scheme for the input speech signal, and the second one (back-end) is a classifier to distinguish between bonafide and spoofed samples. The feature descriptor (front-end) should be capable of effectively capturing the traits of the dynamic vocal tracts of a bonafide speaker. Similarly, the back-end classifier should be able to better learn the distinct traits of bonafide and spoofed speech samples in order to accurately discriminate against spoofed speech. Consequently, voice spoofing detection techniques in early ages have relied on hand-crafted features to identify pitch, duration, and cepstral characteristics of the speech signal [Borrelli et al. 2021; Hemavathi and Kumaraswamy 2021; Phapatanaburi et al. 2019; Xiao et al. 2015]. However, these techniques are often limited in their ability to detect crucial and subtle features of spoofing attacks. In addition, handcrafted feature-based approaches are dependent on relevant expertise and are usually unable to identify subtle characteristics of the speech signal that indicate voice spoofing attacks. Moreover, these techniques may potentially be subject to attacks that target the hand-crafted detecting properties in particular.

In the literature, machine learning solutions have been proposed as a promising approach for detecting voice spoofing attacks. Specifically, there has been significant interest in developing algorithms to detect logical attacks, which exploit the semantic and contextual information of a conversation to deceive the system. In this context, various machine learning techniques have been explored, including SVM, support vector machines, and random forests. Machine learning-based approaches, such as SVM and random forests [Bhangale et al. 2018; Javed et al. 2021; Rahmeni et al. 2020, 2022], have shown promise in detecting voice spoofing attempts by training on large datasets to understand patterns and features of real and imposter voice samples. However, the use of computationally advanced spectrograms can be resource-intensive and computationally expensive, making scaling to larger datasets difficult and increasing the risk of over fitting.

Deep learning-based approaches outperformed traditional machine learning techniques [Chen et al. 2017; Parasu et al. 2020; Tak et al. 2021; Teng et al. 2022; Zhang et al. 2021], but they may require much more training data and computational resources. These techniques may also be prone to overfitting and be incapable of successfully generalizing to new and unknown attacks. Deep learning-based approaches for detecting speech spoofing, such as ResNet [Chen et al. 2017], Res2Net [Tak et al. 2021], ASSIST [Jung et al. 2022], and ASSERT [Lai et al. 2019], also have vulnerabilities. One of the major limitations of these models is their high computational complexity, which can make training and inference prohibitively expensive. This may limit its use in real-world applications requiring quick detection, such as bio-metric authentication systems. Moreover, these models require massive amounts of labeled training data, which might be difficult to get for newer and less common spoofing attacks.

2.0.1 Problem statement and motivation of the proposed solution. Voice spoofing detection is critical in areas such as banking and law enforcement, where malicious actors use techniques to deceive automatic speaker identification systems. Current solutions for detecting voice spoofing rely on either handcrafted or deep learning features. However, combining advanced deep learning architectures

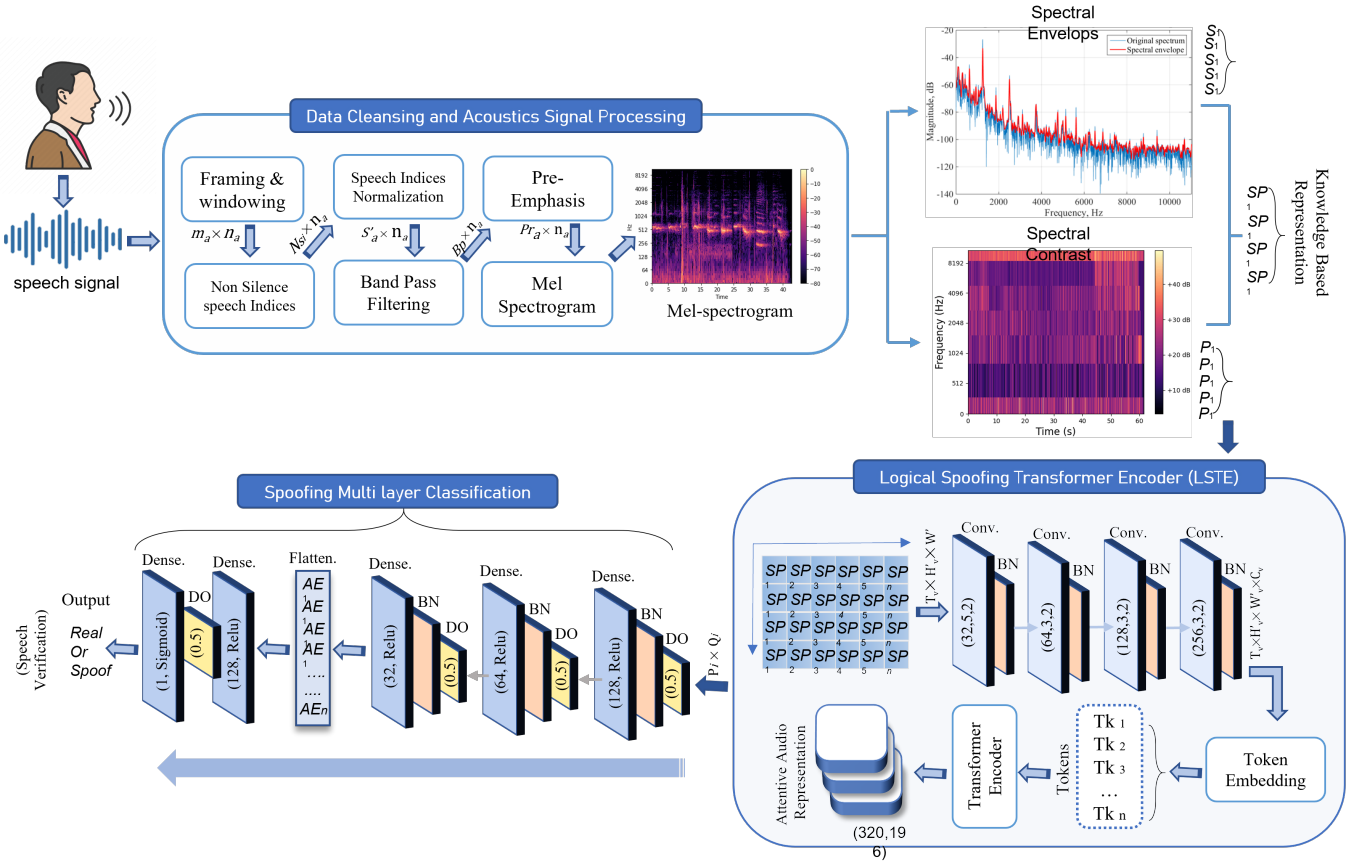

Figure 1: The overall architecture of SpotNet Framework.

such as transformer encoders and decoders with handcrafted features has not been explored well. To address this gap, we propose SpotNet, which uses attention-based attentive spectral and temporal features along with effective handcrafted features for training and classification. We introduce the Logical Spoofing Transformer Encoder (LSTE) to obtain attentive feature maps, and a multi-layer spoofing network to classify them as bona fide or spoofed speech samples. This approach has potential to improve the security of speaker identification systems in sensitive areas such as banking and law enforcement.

3 PROPOSED WORK

This section describes the procedure for developing the voice spoofing detection solution, SpotNet. The proposed framework consists of two folds: In the first fold, data cleansing and essential signal processing to extract reliable spoofing features (FSF) from the raw audio signal, and in the second fold, utilizing the obtained FSF maps as input for the Logical Spoofing Transformer Encoder (LSTE) block, which extracts attentive spectral and temporal characteristics of speech samples using token embedding and transformer encoder. These attentive features are then used for training and classification in the multi-layer spoofing classifier. The architecture of our proposed framework is presented in Fig. 1, and the preprocessing and data cleansing steps are described in section 4.2. The developed multi layer spoofing classifier network with 5 dense layers, followed by batch normalization and dropout, is trained using the extracted attentive features from LSTE to identify authentic and counterfeit audio, potentially improving the security of speaker verification systems.

The Mel-spectrogram, spectral contrast, and spectral envelope are included in the front-end spoofing features (FSF) set to effectively detect voice conversion (VC) and text-to-speech (TTS) logical attacks that manipulate speech signal characteristics. These attacks modify the frequency and spectral properties of speech signals, making it difficult to distinguish between authentic and fake audio. However, the FSF approach captures and analyzes multiple spectral and temporal features, making it highly effective in detecting such attacks. Therefore, the inclusion of these features in the FSF is crucial for achieving accurate and reliable speech spoofing detection. The equation below presents the extraction of these FSF features.

$$M_{m,k} = \sum_{n=0}^{N-1} \frac{1}{N} |x[n] \cdot w[n]| \cdot H_{m,n} \cdot e^{-j2\pi \frac{k \cdot n}{N}} \quad (1)$$

where $M_{m,k}$ is the m-th Mel frequency bin and k-th discrete frequency bin, $H_{m,n}$ is the Mel filterbank, and $w[n]$ is the window

Table 1: Architecture table for the SpotNet model

Layer	Input Shape	Output Shape	Num of Params	Operation
Input	HxWx1	HxWx1	-	-
TokenEmbedding	HxWx1	HxWx2	20,200	Token embedding
TransformerEncoder	HxWx2	HxWx2	168,514	Multi-head attention, feedforward
Conv2D	HxWx2	HxWx16	304	Kernel=3x3, Stride=1, ReLU
BatchNormalization	HxWx16	HxWx16	64	-
Squeeze	HxWx16	HxWx16	-	Squeeze last dimension
MaxPooling2D	HxWx16	$\frac{H}{2}$x$\frac{W}{2}$x16	-	Kernel=2x2, Stride=2, Max pool
Conv2D	$\frac{H}{2}$x250x16	$\frac{H}{2}$x$\frac{W}{2}$x32	2,080	Kernel=3x3, Stride=1, ReLU
BatchNormalization	$\frac{H}{2}$x$\frac{W}{2}$x32	$\frac{H}{2}$x250x32	128	-
MaxPooling2D	$\frac{H}{2}$x$\frac{W}{2}$x32	$\frac{H}{4}$x$\frac{W}{4}$x32	-	Kernel=2x2, Stride=2, Max pool
Conv2D	$\frac{H}{4}$x$\frac{W}{4}$x32	$\frac{H}{4}$x$\frac{W}{4}$x64	8,256	Kernel=3x3, Stride=1, ReLU
BatchNormalization	$\frac{H}{4}$x$\frac{W}{4}$x64	$\frac{H}{4}$x$\frac{W}{4}$x64	256	-
MaxPooling2D	$\frac{H}{4}$x$\frac{W}{4}$x64	$\frac{H}{6}$x$\frac{W}{6}$x64	-	Kernel=2x2, Stride=2, Max pool
Conv2D	$\frac{H}{6}$x$\frac{W}{6}$x64	$\frac{H}{6}$x$\frac{W}{6}$x128	32,896	Kernel=3x3, Stride=1, ReLU
BatchNormalization	$\frac{H}{6}$x$\frac{W}{6}$x128	$\frac{H}{6}$x$\frac{W}{6}$x128	512	-
MaxPooling2D	$\frac{H}{6}$x$\frac{W}{6}$x128	$\frac{H}{8}$x$\frac{W}{8}$x128	-	Kernel=2x2, Stride=2, Max pool
Flatten	$\frac{H}{8}$x$\frac{W}{8}$x128	12,288	-	Flatten to 1D
Dense	12,288	32	393,248	ReLU
Dropout	32	32	-	Dropout (p=0.5)
Dense	32	1	33	Sigmoid
Total Parameters			**603,361**	

function. The Spectral Contrast and Spectral Envelope are computed using the following equations:

$$C_{m,k} = \log\left(\frac{\sum_{i=1}^{n} \omega_i |M_{m,k+\omega_i}|^2}{\sum_{i=1}^{n} \omega_i |M_{m,k-\omega_i}|^2}\right) \tag{2}$$

$$E_m = \sum_{k=0}^{K} |M_{m,k}| \cdot w[k] \tag{3}$$

where $C_{m,k}$ is the m-th contrast coefficient for the k-th discrete frequency bin, E_m is the m-th envelope coefficient, n is the number of frequency bins used to calculate contrast, ω_i is the weight of the i-th frequency bin, and $w[k]$ is the weighting function.

3.1 Logical Spoofing Transformer Encoder

The Logical Spoofing Transformer Encoder (LSTE) is utilized to extract deep attentive features from the FSF before classification. LSTE is a advanced deep learning model that captures attentive spectral and temporal traits of speech signals and generates a compact and informative representation of the audio. The transformer encoder block is presented in Table 2. By extracting attentive features, we reduce the dimensionality of the feature space, identify the most relevant features, and enhance the robustness of the classifier. These features aid in distinguishing between real and spoofed audio and reducing the impact of irrelevant or noisy features.

3.2 Construction of Proposed Classifier

A multi-layer classification network is designed to classify speech signals into genuine or fake categories using the obtained attentive spectral temporal features. The network consists of five dense layers followed by batch normalization and dropout, which can effectively distinguish real and spoof audio. The classifier is based on a dense network, which can learn complex and non-linear relationships between input features and output classes, improving the classification accuracy. The architecture of the classifier can be found in Figure 1 and Table 1.

4 EXPERIMENTAL RESULTS

This section presents the experimental setup and methodology employed to evaluate the effectiveness of the proposed SpotNet approach in detecting voice spoofing. It includes information about the dataset, evaluation metrics, and hyper-parameters used in the modeling and training process of the SpotNet solution.

4.1 Dataset and Metrics

The effectiveness of the proposed SpoTNet method for detecting voice spoofing attacks was evaluated using the ASVspoof2019 dataset, specifically its LA subpart. The ASVspoof2019-LA dataset is widely used as a standard dataset for testing speaker verification systems and comprises 22,800 spoofed and 2,580 real speech samples, featuring various TTS and VC spoofing technologies such as neural waveform models and vocoders. The training subset of the dataset was utilized for training the model, while the development subset was used for validation purposes. The evaluation subset was then used to assess the performance of the model using a range of evaluation metrics such as EER, min-tDCF, precision, recall, F1-score, and accuracy. The EER and min-tDCF metrics, which are commonly used to evaluate voice spoofing detection systems, were used to compare the performance of the proposed method against

Table 2: Transformer Encoder architecture. In this table L refers to the length of the input sequence and D refers to the dimensional of the embedding vectors. The total number of parameters in the model is $8D^2 + 4D$.

Layer	Input Shape	Output Shape	Operation	Num of Params
Input	L	$L \times D$	-	-
Positional Encoding	$L \times D$	$L \times D$	-	0
Multi-Head Attention	$L \times D$	$L \times D$	$\text{Attention}(Q, K, V)$	$3KD^2$
Addition	$L \times D$	$L \times D$	$x + \text{Attention}(Q, K, V)$	0
Layer Normalization	$L \times D$	$L \times D$	-	$2D$
Position-wise Feedforward	$L \times D$	$L \times D$	$\text{ReLU}(xW_1 + b_1)W_2 + b_2$	$4D^2$
Addition	$L \times D$	$L \times D$	$x + \text{ReLU}(xW_1 + b_1)W_2 + b_2$	0
Layer Normalization	$L \times D$	$L \times D$	-	$2D$

Algorithm 1: Speech Sample FSF embedding extraction

Input: Raw audio waveform x, Sampling rate f_s
Output: Speaker embedding z

$w \leftarrow \text{Window}(x)$ // windowing and Framing to x
$F \leftarrow \text{Frame}(w)$ $S \leftarrow \text{NonSilenceIndices}(F)$ // Retain non-silent content
$F_S \leftarrow F(S, :)$ $F_N \leftarrow \text{Normalize}(F_S)$ // Normalize mean and unit variance
$F_B \leftarrow \text{BandPassFilter}(F_N, f_n)$ // band pass filtering for frequencies between 20 Hz and 8 kHz
$F_{PR} \leftarrow \text{PreEmphasisFilter}(F_N)$ // pre-emphasis filter
$M \leftarrow \text{MelSpectrogram}(F_{PR}, f_{pr})$ // Compute Mel-scaled spectrogram with 40 bins
$C \leftarrow \text{SpectralContrast}(M)$ // Compute spectral contrast features
$E \leftarrow \text{SpectralEnvelope}(M)$ // Compute spectral envelope features
$X \leftarrow \text{Concatenate}(M, C, E)$ // Concatenate the feature matrices along the channel axis
$X^{'} \leftarrow \text{PadOrTruncate}(X, (48, 501, 3))$ // Pad or truncate the feature tensor
$z \leftarrow \text{LSTE}(X)$

state-of-the-art techniques. Additionally, other statistical parameters such as precision, recall, F1-score, and accuracy were employed to evaluate the overall performance of the proposed method.

4.2 Data cleansing and pre-processing

The proposed work aims to enhance the efficiency of the SpotNet model by integrating five preprocessing methods. The first step is to divide the raw audio stream into short overlapping frames using the windowing and framing process, as shown in the equation:

$$y(n) = w(n) \cdot x(n) \tag{4}$$

where $x(n)$ is the input audio signal, $w(n)$ is the window function, and $y(n)$ is the windowed output signal. Then, non-silence indices retrieval technique is employed to eliminate the silent segments from the audio signal. Silence intervals in audio recordings contain no speech information and can lead to false positives or false negatives in audio signal analysis [Sahoo and Patra 2014]. Thus, we extract the non-silence indices from the speech signal using the following equation:

$$n_i = \arg\max_n \frac{\sum_{m=n-T+1}^{n} x_m^2}{T} > \epsilon \tag{5}$$

where $x(n)$ is the input audio signal, T is the frame size, ϵ is the energy threshold, and n_i is the index of the first non-silent frame. Next, the normalization technique is applied to normalize the amplitude of the audio signal. The equation used for normalization is given by:

$$y(n) = \frac{x(n)}{\max_n |x(n)|} \tag{6}$$

where $x(n)$ is the input audio signal and $y(n)$ is the normalized output signal. Following this, the bandpass filtering technique is applied to remove unwanted frequency components from the audio signal using the equation:

$$y(n) = x(n) * h(n) \tag{7}$$

where $x(n)$ is the input audio signal, $h(n)$ is the bandpass filter, and $*$ denotes convolution. Finally, the pre-emphasis filtering technique is applied to boost the high frequency components of the audio signal. This technique can be represented by the equation:

$$y(n) = x(n) - \alpha x(n-1) \tag{8}$$

where $x(n)$ is the input audio signal, $y(n)$ is the pre-emphasized output signal, and α is the pre-emphasis coefficient. After applying the aforementioned preprocessing techniques, the front-end spoofing feature set extraction is performed that explained in the next subsection below.

4.3 Front-end Spoofing Feature set (FSF) Extraction

This work aims to extract effective features from voice samples using robust feature extraction techniques. The audio signals are segmented into small segments with a window size of w_n of 0.025s and a step size of h = 0.01s. The Fourier transform FFT with a window function is applied on these segments to transform the input signals into the frequency domain. The resulting power spectrum is mapped onto the Mel scale to obtain Mel spectrogram components. Mel spectrogram features are extracted using 40 mel filters and adjusted the lowest and highest frequencies. Spectral contrast

Table 3: Statistics of ASVspoof2019-LA Dataset.

Subset	# of Instances	Spoofing Algorithms	Waveform Generator
Training	25,380	–	–
Development	24,844	A01-A06	WaveNet, WORLD, Waveform, Spectral Filtering
Evaluation	71,237	–	STRAIGHT, +OLA, Vocaine

Table 4: Statistics of Clonning ALgorithms of ASVspoof2019-LA Dataset.

Subset	Training	Development	Spoofing System	Algorithm	Waveform Generator
A01	3,800	3,706	TTS	Neural Waveform Model	WaveNet
A02	3,800	3,706	TTS	Vocoder	WORLD
A03	3,800	3,706	TTS	Vocoder	WORLD
A04	3,800	3,706	TTS	Waveform Concatenation	Waveform Concatenation
A05	3,800	3,706	VC	Vocoderl	WORLD
A06	3,800	3,706	VC	Spectral filtering	Spectral filtering + OLA

and spectral envelope features are also extracted. Spectral contrast evaluates the difference in spectral energy between neighboring frequency bands, while the spectral envelope reflects the level of spectral flatness inside each frame. These features are concatenated into a single feature vector and fed into the LSTE block to extract attentive spectral temporal features with an attention mechanism. The dimension of the resulting feature vector is (48, t), where t is the maximum number of frames.

4.4 Attentive Feature Extraction from Logical Spoofing Transformer Encoder and Classification

The proposed LSTE architecture extracts attentive spectral temporal features from voice samples for voice spoofing detection. The LSTE block comprises token embedding and a transformer encoder with multiple stacked transformer blocks. The output of the transformer encoder is fed into a dense architecture-based spoofing multi-layer classification network for classification. The proposed LSTE architecture is shown in Table 2, while the architecture details of the multi-layer classification network are presented in Table 1.

4.5 Hyper parameter Modeling and Training

We perform the training of the model on four NVIDIA Tesla V100 16G GPUs, 192 GB of Memory, and 48 CPU cores running at 2.10 GHz. The model takes input of shape ($HxWxl$) which is (48, 501, 1) and uses Token Embedding layer with a vocabulary size of V=100, maximum sequence length of N=100, and a hidden size of h=32. The Transformer Encoder layer has a hidden size of (MxN) is 4, 4 attention heads, feed-forward dimension of 128, and a dropout rate of d=0.1. The CNN layer has four 2D convolutional layers with f_1=16, f_2= 32, f_3= 64, and f_4= 128 filters, respectively, and kernel sizes of (k_1) = (3,3), (k_2) = (2,2), (k_3) = (2,2), and (k_4) = (2,2), respectively. The model uses a batch size of 32 and a binary cross-entropy loss function with metrics including AUC and binary accuracy. The optimizer used is Adam with an initial learning rate of lr 0.001 and a learning rate schedule with an exponential decay rate of 0.3 every 4800 steps.

4.6 The Performance Analysis of the Proposed SpotNet

We evaluated the proposed system's performance on the ASVspoof19-LA dataset and present the results in Table 6. The system utilized attentive spectral and temporal features, achieving an EER of 0.95% and an AUC of 94.11%. We assessed the effectiveness of the system against each voice cloning algorithm present in the corpus (ranging from A01-A06) and reported the results in Table 7. Our model obtained an EER of 0.08%, 0.11%, 0.10%, 0.23%, 0.09%, and 0.21% for A01-A06 spoofing algorithms, respectively. These results demonstrate that the proposed system performed optimally against each voice cloning algorithm. Additionally, we evaluated the system's performance against an unseen evaluation spoofing algorithm and report the results in Table 8. The model was effective against most types of spoofing attacks, achieving a lower EER and higher accuracy. While the EER for spoofing algorithms A17–A19 was slightly higher than for other spoofing algorithms, the model achieved an EER of less than 1.0% for the remaining 10 unseen and 6 seen spoofing algorithms, indicating its ability to generalize well to unknown and unseen voice spoofing attacks.

4.7 Comparative Analysis of the Proposed SpotNet and SOTA feature fusion based techniques

In this research, we compared the performance of the proposed SpotNet solution against eight recent studies that utilized feature fusion-based techniques. The results of the comparison are presented in Table 5 and Figure 2, which shows that the proposed SpotNet solution achieves comparable performance with other feature fusion-based techniques. The proposed method takes Mel-spec, spec-contrast, and spec-envelope-based robust features as input, extracts attentive features from them, and achieves a min-tDCF of 0.045 and an EER of 0.95%. These results are better than most

Table 5: Comparative Analysis of proposed method with Frontend feature fusion studies.

Study	Fusion System		Performance	
	Classifier	Frontend Features	t-DCF	EER(%)
2019 [Alzantot et al. 2019]	ResNet	MFCC, Spec, CQCC	0.157	6.02
2019 [Lavrentyeva et al. 2019]	LCNN	LFCC,CQT,FFT, LFCC+CMVN	0.051	1.84
2020 [Wang et al. 2020]	DenseNet	spec,LFCC	0.047	1.98
2021 [Luo et al. 2021]	Capsule	LFCC, STFT-gram	0.033	1.07
2021 [Li et al. 2021]	SE-ResNet50	Spec,LFCC,CQT	0.045	1.89
2021 [Zhang12 et al. 2021]	SENet	dual-band of FFT	0.050	1.56
2022 [Wei et al. 2022]	GMM,LCNN	LFCC, CQCC, RLFCC	0.074	2.57
2022 [Cui et al. 2022]	scDenseNet	Spec,LFCC,ARS	0.029	1.01
2022 [Cui et al. 2022]	scDenseNet	SpecL,LFCC,ARS	0.042	0.98
Proposed	SpotNet	mel-spec, contrast, Spec-Env	0.045	0.95

Table 6: Performance on ASVspoof19-LA Dataset.

Model	EER %	min-tDCF	Accuracy %	Precision %	Recall %	F1-Score %	Auc %
SpotNet	0.95	0.045	93.91	93.32	97.22	95.25	94.11

Table 7: Performance analysis on voice clonning algorithms of ASVspoof2019.

Algorithm	EER%	Acc%	Precision%	Recall%	F1-score%
A01	0.08	97.5	94.29	93.20	94.12
A02	0.11	94.6	92.36	92.55	92.42
A03	0.10	98.3	93.63	94.25	93.66
A04	0.23	92.6	90.90	89.56	91.25
A05	0.09	95.3	94.88	94.23	95.36
A06	0.21	0.89	91.32	91.33	90.32

Table 8: Performance analysis on unseen voice spoofing algorithms of ASVspoof2019.

Algorithm	EER%	Acc%	Precision%	Recall%	F1-score%
A07	0.40	97.50	97.29	93.20	95.50
A08	0.39	98.67	98.36	92.55	95.42
A09	6.10	93.35	82.69	80.25	79.66
A10	0.42	95.60	97.90	92.40	93.25
A11	0.42	95.30	95.88	91.23	93.36
A12	0.40	97.89	93.32	92.33	91.17
A13	0.39	96.50	98.29	93.80	94.78
A14	0.37	96.63	98.36	94.55	95.18
A15	0.40	98.34	97.63	95.25	95.66
A16	0.45	94.62	93.90	91.56	92.25
A17	25.31	75.39	65.88	60.23	63.36
A18	40.23	59.87	55.90	53.56	52.56
A19	44.09	55.93	52.88	40.23	45.36

state-of-the-art comparative methods and are closest to those of the scDeneNet model [Cui et al. 2022].

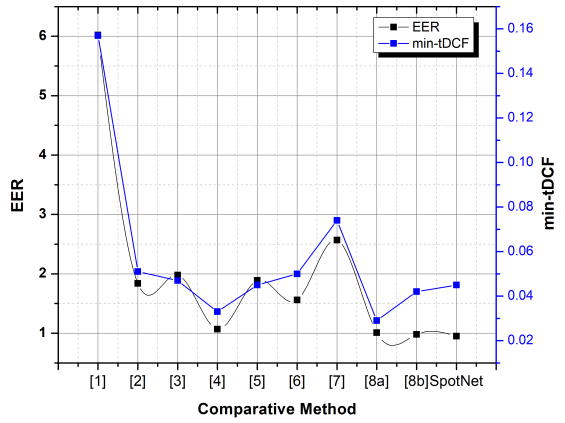

Figure 2: The overall architecture of SpotNet Framework.

Although the min-tDCF of the proposed solution is slightly higher than that of [Cui et al. 2022] and [Luo et al. 2021], it is important to note that our primary goal was to achieve a lower EER without performing speaker verification. This is important because in some scenarios, such as in security and authentication systems, detecting spoofing attacks is more critical than verifying the speaker's identity. Thus, a slightly higher t-DCF is acceptable in this case. Our study also reveals that the countermeasure containing spectrogram features concatenated with other spectral features, such as LFCC, CQCC, CQT, contrast, and envelope, performed optimally within the range of classifiers shown in Table 5. This highlights the effectiveness of spectrogram features in detecting logical access attacks and justifies their use in the proposed method for spoofing detection.

Table 9: Performance analysis with different combination of the frontend features.

Performance	Mel-Spec	Spec-contrast	Spec-Envelop	mel-Spec+Contrast	Mel-Spec+Env	Mel-Spec+Env+Cont
t-DCF/EER	0.109/5.49	0.221/7.14	0.262/10.87	0.101/5.65	0.127/4.84	0.045/0.95

4.8 Ablation Study

In this research, we conducted an ablation analysis using the SpotNet model to identify the optimal feature combination for detecting spoofing attacks. We evaluated the performance of the model on standalone Mel-spectrogram, spectral envelop, and spectral contrast features, as well as all possible combinations of the three features. The results of the study, presented in Table 9, revealed that the EER for Mel-spec, spec-contrast, spec-envelope, Mel-spec and contrast, and Mel-spec with envelope feature sets were 5.49%, 7.14%, 10.87%, 5.65%, and 4.84%, respectively.

Our findings suggest that the best performance was achieved when all three features were combined, resulting in an EER of 0.95% and a min-tDCF of 0.042. We also tested the SpotNet model with a feature shape of 25,400, where we computed and classified 20 Mel-spec, 4 spectral contrasts, and 1 envelope using the lighter SpotNet-model. The results were comparable to those obtained using the optimal feature combination, with good performance observed for both EER and min-tDCF. In conclusion, our ablation study demonstrated that combining the Mel-spectrogram, spectral envelope, and spectral contrast features is the most effective approach for detecting spoofing attacks using the SpotNet model.

5 CONCLUSION AND FUTURE WORK

In this study, we introduced a novel method for detecting spoofing attacks in automated speaker verification systems using the proposed SpoTNet model. Our experimental results demonstrate that the proposed approach, which integrates handcrafted features with attention-based features, outperforms state-of-the-art methods on the ASVspoof 2019 dataset. However, it is worth noting that our approach need for a large amount of training data to train the transformer encoder. To address this limitation, our future research will focus on exploring other methods, such as few-shot learning and utilizing raw audios, to reduce the data dependencies and hand crafted and spectrogram based computational complexities. Furthermore, we plan to extend this work to other types of voice spoofing attack detection and develop solutions that focus exclusively on real speech samples, capable of detecting unseen and novel attacks and generalizability of the countermeasures.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (NSF) under award number 1815724 and Michigan Transnationals Research and Commercialization (MTRAC), Advanced Computing Technologies (ACT) award number 292883. Any opinions, findings, and conclusions in this material do not necessarily reflect the views of the NSF and MTRAC ACT. document.

REFERENCES

Moustafa Alzantot, Ziqi Wang, and Mani B Srivastava. 2019. Deep residual neural networks for audio spoofing detection. *arXiv preprint arXiv:1907.00501* (2019).

Kishor B Bhangale, Prashant Titare, Raosaheb Pawar, and Sagar Bhavsar. 2018. Synthetic speech spoofing detection using MFCC and radial basis function SVM. *IOSR J. Eng.(IOSRJEN)* 8, 6 (2018), 55–62.

Clara Borrelli, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro. 2021. Synthetic speech detection through short-term and long-term prediction traces. *EURASIP Journal on Information Security* 2021, 1 (2021), 1–14.

Zhuxin Chen, Zhifeng Xie, Weibin Zhang, and Xiangmin Xu. 2017. ResNet and Model Fusion for Automatic Spoofing Detection.. In *Interspeech*. 102–106.

Sanshuai Cui, Bingyuan Huang, Jiwu Huang, and Xiangui Kang. 2022. Synthetic Speech Detection Based on Local Autoregression and Variance Statistics. *IEEE Signal Processing Letters* 29 (2022), 1462–1466. https://doi.org/10.1109/LSP.2022.3183951

R Hemavathi and R Kumaraswamy. 2021. Voice conversion spoofing detection by exploring artifacts estimates. *Multimedia Tools and Applications* 80 (2021), 23561–23580.

Bingyuan Huang, Sanshuai Cui, Jiwu Huang, and Xiangui Kang. 2023. Discriminative Frequency Information Learning for End-to-End Speech Anti-Spoofing. *IEEE Signal Processing Letters* 30 (2023), 185–189.

Ali Javed, Khalid Mahmood Malik, Aun Irtaza, and Hafiz Malik. 2021. Towards protecting cyber-physical and IoT systems from single-and multi-order voice spoofing attacks. *Applied Acoustics* 183 (2021), 108283.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2022. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6367–6371.

Shrikrishna V Kulkarni and Shrikrishna A Khaparde. 2017. *Transformer engineering: design, technology, and diagnostics*. CRC press.

Cheng-I Lai, Nanxin Chen, Jesús Villalba, and Najim Dehak. 2019. ASSERT: Anti-spoofing with squeeze-excitation and residual networks. *arXiv preprint arXiv:1904.01120* (2019).

Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov. 2019. STC antispoofing systems for the ASVspoof2019 challenge. *arXiv preprint arXiv:1904.05576* (2019).

Xu Li, Na Li, Chao Weng, Xunying Liu, Dan Su, Dong Yu, and Helen Meng. 2021. Replay and synthetic speech detection with res2net architecture. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 6354–6358.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3202–3211.

Anwei Luo, Enlei Li, Yongliang Liu, Xiangui Kang, and Z Jane Wang. 2021. A capsule network based approach for detection of audio spoofing attacks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6359–6363.

Prasanth Parasu, Julien Epps, Kaavya Sriskandaraja, and Gajan Suthokumar. 2020. Investigating Light-ResNet Architecture for Spoofing Detection Under Mismatched Conditions.. In *INTERSPEECH*. 1111–1115.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. 2018. Image transformer. In *International conference on machine learning*. PMLR, 4055–4064.

Khomdet Phapatanaburi, Longbiao Wang, Seiichi Nakagawa, and Masahiro Iwahashi. 2019. Replay attack detection using linear prediction analysis-based relative phase features. *IEEE Access* 7 (2019), 183614–183625.

Raoudha Rahmeni, Anis Ben Aicha, and Yassine Ben Ayed. 2020. Acoustic features exploration and examination for voice spoofing counter measures with boosting machine learning techniques. *Procedia Computer Science* 176 (2020), 1073–1082.

Raoudha Rahmeni, Anis Ben Aicha, and Yassine Ben Ayed. 2022. Voice spoofing detection based on acoustic and glottal flow features using conventional machine learning techniques. *Multimedia Tools and Applications* 81, 22 (2022), 31443–31467.

Tushar Ranjan Sahoo and Sabyasachi Patra. 2014. Silence removal and endpoint detection of speech signal for text independent speaker identification. *International Journal of Image, Graphics and Signal Processing* 6, 6 (2014), 27.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. 2021. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6369–6373.

Zhongwei Teng, Quchen Fu, Jules White, Maria E Powell, and Douglas C Schmidt. 2022. ARawNet: A Lightweight Solution for Leveraging Raw Waveforms in Spoof Speech Detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE,

692–698.

Zheng Wang, Sanshuai Cui, Xiangui Kang, Wei Sun, and Zhonghua Li. 2020. Densely connected convolutional network for audio spoofing detection. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 1352–1360.

Linqiang Wei, Yanhua Long, Haoran Wei, and Yijie Li. 2022. New acoustic features for synthetic and replay spoofing attack detection. *Symmetry* 14, 2 (2022), 274.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.

Xiong Xiao, Xiaohai Tian, Steven Du, Haihua Xu, Engsiong Chng, and Haizhou Li. 2015. Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge.. In *Interspeech*. 2052–2056.

You Zhang, Fei Jiang, and Zhiyao Duan. 2021. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters* 28 (2021), 937–941.

Yuxiang Zhang12, Wenchao Wang12, and Pengyuan Zhang12. 2021. The effect of silence and dual-band fusion in anti-spoofing system. (2021).