ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447500

FRAME-TO-UTTERANCE CONVERGENCE: A SPECTRA-TEMPORAL APPROACH FOR UNIFIED SPOOFING DETECTION

Awais Khan[1], *Khalid Mahmood Malik*[1,2], *Shah Nawaz*[3]

[1]Department of Computer Science and Engineering, Oakland University, Rochester, Michigan, USA
[2]College of Innovation and Technology, University of Michigan-Flint, Michigan, USA
[3]Johannes Kepler University, Linz, Austria

ABSTRACT

Voice spoofing attacks pose a significant threat to automated speaker verification systems. Existing anti-spoofing methods often simulate specific attack types, such as synthetic or replay attacks. However, in real-world scenarios, the countermeasures are unaware of the generation schema of the attack, necessitating a unified solution. Current unified solutions struggle to detect spoofing artefacts, especially with recent spoofing mechanisms. For instance, the spoofing algorithms inject spectral or temporal anomalies, which are challenging to identify. To this end, we present a spectra-temporal fusion leveraging frame-level and utterance-level coefficients. We introduce a novel local spectral deviation coefficient (SDC) for frame-level inconsistencies and employ a bi-LSTM-based network for sequential temporal coefficients (STC), which capture utterance-level artifacts. Our spectra-temporal fusion strategy combines these coefficients, and an auto-encoder generates spectra-temporal deviated coefficients (STDC) to enhance robustness. Our proposed approach addresses multiple spoofing categories, including synthetic, replay, and partial deepfake attacks. Extensive evaluation on diverse datasets (ASVspoof2019, ASVspoof2021, VSDC, partial spoofs, and in-the-wild deepfakes) demonstrated its robustness for a wide range of voice applications.

Index Terms— Voice Spoofing Detection, Spectral Temporal, Audio Deepfake Detection, Unified spoofing detection

1. INTRODUCTION

Voice authentication methods are mainstream solutions for identity verification systems, but the increasing prevalence of voice spoofing, including logical, physical, and deepfake attacks, poses a significant threat to their effectiveness [1]. Existing methods often focus on mitigating individuals or a subset of these attacks. For example, recent research [2] highlights this vulnerability, particularly for partial and full deepfakes. While existing systems effectively detect replay and synthetic speech, they struggle to identify partial deepfakes, as shown in Table 1. The results show that the countermeasure demonstrates a substantial drop in performance when evaluated on partially spoofed samples. Even training specifically on partial spoofs did not fully address the performance issue. In the speech spectral analysis shown in Fig.1, a partial spoofing spectrogram differs from spectrograms of replay or synthetic speeches and showing heterogeneous spectral artifacts. This may lead to significant performance deterioration of existing spoofing countermeasures. This challenge persists even when these systems are trained specifically on dataset of partial spoofs, underscoring the necessity for a solution proficient in detecting temporal disparities in partial deepfake scenarios.

Table 1: An architecture from the top 5 performers of the ASV challenge [2] is evaluated in terms of generalizability using ASVspoof2019-LA and Partialspoof2021 datasets. (Lower is better).

		ASV		PSF	
	Train	Dev.	Eval.	Dev.	Eval.
EER(%)	ASV	0.21	2.65	9.59 ↑	15.96 ↑
	PSF	4.28 ↑	5.38 ↑	3.68	6.19 ↑
min-tDCF	ASV	0.006	0.064	0.185 ↑	0.300 ↑
	PSF	0.115	0.171	0.100	0.164

Previous anti-spoofing techniques aimed at preventing either physical or logical attacks [1, 3, 4, 5]. However, recent approaches focus on developing a unified solution based on utterance-level features capable of detecting both physical and logical attacks (LA) [6, 7, 8]. Despite this, these unified solutions tend to exhibit bias towards either detecting logical or physical attacks (PA), highlighting the necessity for an impartial unified solution. Additionally, other research has delved into the detection of partial and fully deepfake attacks in a unified solution based on segment-level features [6, 9]. However, these methods often fall short in identifying physical attacks. Therefore, addressing both full and partial deepfake attacks requires a comprehensive approach that considers both segment-level and utterance-level artifacts.

To address these challenges, we propose a spectra-temporal approach involving the extraction of frame-oriented spectral deviated coefficients (SDC) and utterance-oriented sequential temporal coefficients (STC) using a Bidirectional Long Short-Term Memory (Bi-LSTM) network. The rationale behind STC and SDC lies in analyzing spectral inconsistencies within distinct frequency ranges. Replay

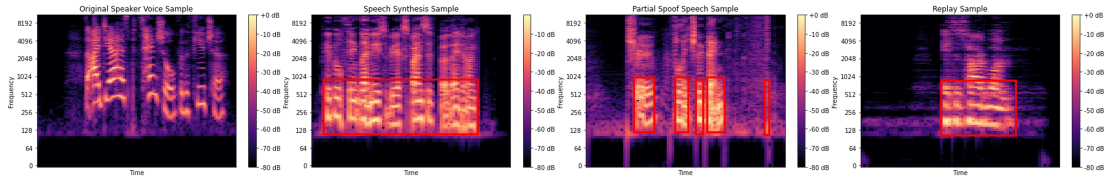

Fig. 1: Spectrogram comparison of bona fide (first-left), fully synthesized (second), partially deep fake (third), and replay (fourth) speech samples.

attacks, involving the mic-speak-mic process, exhibit pronounced non-linearities and microphonic distortion in lower frequency bands, constrained by the recording or replaying microphones. Conversely, synthetic speech generated by AI algorithms leads to spectral manipulation primarily in higher frequency ranges due to the absence of a real-speaker vocal frequency representation. Additionally, STC tackles temporal inconsistencies in partial spoofing. These components collectively capture intricate patterns at both the utterance and frame levels, forming a strong foundation for our unified approach.

The main contributions of this paper are as follows: 1) We introduce a spectra-temporal-based unified method for the detection of different voice spoofing categories. 2) We proposed spectral deviated coefficients for segment-level artifact extraction and employed a bi-LSTM network to capture sequential temporal artifacts within speech signals. Through rigorous experimentation using diverse datasets, we demonstrate the effectiveness of the proposed method. To the best of our knowledge, this is the first ever attempt to tackle four different types of voice spoofing with a single system.

2. PROPOSED METHOD

The proposed method is divided into three sections, as shown in Fig. 2. It consists of Spectral Deviated Coefficients (SDC), Sequential Temporal Coefficients (STC), and Spectra-Temporal Deviation Coefficients (STDC). These sections collectively form a unified method for the reliable detection of voice spoofing.

2.1. Spectral Deviated Coefficients (SDC)

We used the raw input speech signal $s(t)$ to extract SDC, consisting of both higher and lower frequencies across various time frames:

$$s(t) = h * sin(2\pi f_1 t) + l * sin(2\pi f_2 t) \tag{1}$$

where h and l represent the amplitudes of frequencies, and f_1 and f_2 denote the higher and lower frequencies, respectively. Next, we use Hamming windows, which minimizes the spectral leakage by tapering frame edges and preventing abrupt truncation:

$$w[n] = \alpha - \beta \cdot \cos\left(\frac{2\pi n}{N-1}\right) \tag{2}$$

$$y[n] = s[t] \cdot w[n] \tag{3}$$

where $s[t]$ denotes the input signal, $w[n]$ represents the Hamming window with a size of N, and α and β are the window center and edge coefficients, respectively. The resulting segmented signal, after applying windowing and framing, is denoted as $y[n]$. Next, we transform the obtained $y[n]$ to the frequency spectra using a log-Mel spectrogram and fast Fourier transform (FFT) with the following parameters (hop length = 512, mels = 128, fft = 2048) as follows:

$$S[mk] = \log\left(1 + \sum_{n=0}^{N-1} |X[n]|^2 \cdot H_m[k, f_n]\right) \tag{4}$$

where $S[mk]$ represents the log-Mel spectrogram at Mel frequency m and frame k, $X[n]$ stands for the Short-Time Fourier Transform (STFT) at time n, and $H_m[k, f_n]$ represents the Mel filterbank at frequency f_n corresponding to Mel frequency m. The obtained log-transformed Mel spectrogram is then subjected to the Local Deviated Pattern (LDP) operator, which captures the local higher and lower frequency spectrum as follows:

$$LDP(S_{mk(c)}, S_{mk(n)}, \mu_t) = \begin{cases} 1 & S_{mk(n)} \geq S_{mk(c)} + \mu_t, \\ -1 & S_{mk(n)} \leq S_{mk(c)} - \mu_t, \\ 0 & S_{mk(c)} - \mu_t \leq S_{mk(n)}, \\ 0 & S_{mk(n)} \leq S_{mk(c)} + \mu_t \end{cases} \tag{5}$$

where $LDP(S_{mk(c)}, S_{mk(n)}, \mu_t)$ represents the Local Deviated Pattern at position (c, n) with $S_{mk(c)}$ and $S_{mk(n)}$ representing the central and neighboring window values, and μ_t refers to the central tendency average of the window. We determine the conditioning threshold by considering both $S_{mk(c)}$ and μ_t, rather than relying solely on the central window value. It enhances the extraction of LDP features by capturing deviations from the central value, revealing patterns indicative of underlying acoustic traits.

To efficiently handle spatial frequencies, we separately process the higher and lower frequencies of $S[mk]$. The LDP employs triplicate conditions to extract both higher and lower patterns. These patterns are further categorized into two sets: local higher spectra (LHS) and local lower spectra (LLS). Before computing LHS and LLS, we transform negative values into positive ones, as shown in Eqs. 6 and 7. For LHS, we convert all '-1' values to '0' while leaving the other values unchanged, as described in Eq. 6 This results in a set of positive higher-order patterns in $S[mk]$. Similarly, LLS patterns are derived by replacing '1' with '0' and '-1' with '1' in

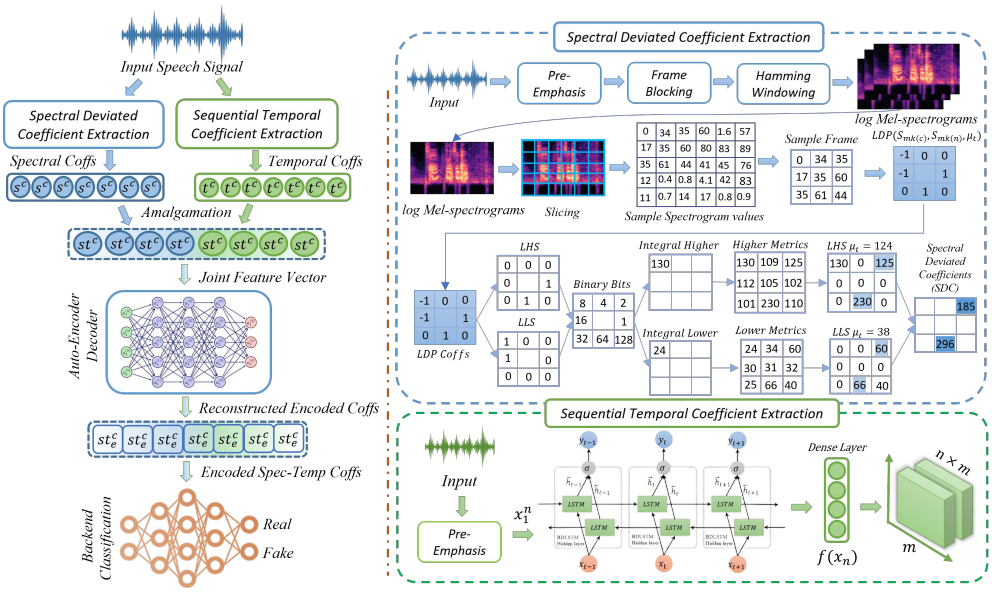

Fig. 2: Architectural diagram of the proposed solution (left). The right upper subset in blue dotted line represent the extraction mechanism of frame level Spectral Deviated Coefficients. The right lower subset in green presents the extraction mechanism of utterance level Sequential Temporal Coefficients.

$LDP(S_mk(c), S_mk(n), mu_t)$ as follows:

$$LHS = LDP(S_{mk(c)}, S_{mk(n)}, \mu_t) = -1 \rightarrow 0 \quad (6)$$

$$LLS = \begin{cases} LDP(S_{mk(c)}, S_{mk(n)}, \mu_t) = 1 \rightarrow 0 \\ LDP(S_{mk(c)}, S_{mk(n)}, \mu_t) = -1 \rightarrow 1 \end{cases} \quad (7)$$

The binary bit streams, denoted as LHS and LLS, are converted into decimal values through a bit extraction process. We begin by extracting bits from the eastern direction and proceed in a counter-clockwise manner to obtain the decimal equivalents as shown in the equation below:

$$HL_{(int)} = \sum_{i=0}^{K-1} HL(C_{rn}) \times 2^{i-1} \quad (8)$$

where HL denotes the higher and lower coefficients obtained from Eqs. 6 and 7, C_{rn} represents the right neighbour at each position, and K is the total number of bits. Next, we extract deviated tendency patterns from the obtained $HL_{(int)}$ to ensure the presence of spectral artifacts in both lower and higher spectral coefficients. Later, we only extract coefficients that exist in both higher and lower coefficients and neglect the rest of the values. We perform this task in a two-step process. First, we compute the mean vector of both higher and lower integrals separately as follows:

$$MV_{(\delta)} = \frac{1}{n} \sum_{i=1}^{n} HL_{(int)} \quad (9)$$

where $MV_{(\delta)}$ refers to the mean vector from higher and lower integrals $HL_{(int)}$. Next, we compute the central tendency vector from the obtained mean vectors $HT_{(\delta)}$ as follows:

$$CTV_{(\delta)} = \frac{1}{n} \sum_{i=1}^{n} MV_{(\delta)} \quad (10)$$

where $CTV_{(\delta)}$ denotes the central tendency mean value from the obtained mean vectors in Eq. 9. By calculating the mean from the mean vectors, we confirm the presence of higher frequencies in both higher and lower integrals, combining them into a single optimal SDC. We retained values that are higher than their mean values and added them to derive the optimal robust spectral features, as demonstrated in Eq. 11.

$$SDC_{(coff)} = [HL_{(int)} > CTV_{(\delta)}] \quad (11)$$

where $SDC_{(coff)}$ represents the spectral deviated coefficients. Finally, a discrete Fourier transform (DFT) is applied to the LDP-transformed $SDC_{(coff)}$ coefficients to obtain robust 128D spectral features. The upper right side of Fig. 2 shows the extraction of SDC patterns.

2.2. Sequential Temporal Coefficients (STC)

We employed a bidirectional long-short-term memory (Bi-LSTM) network to extract sequence-based utterance-level features. Bi-LSTM's bidirectional processing, unlike traditional LSTMs, considers both backward and forward context, enhancing complex temporal relationships. In this work, a two-layer Bi-LSTM configuration [10] was employed to improve temporal feature extraction, yielding 128-dimensional temporal features.

2.3. Spectra-Temporal Deviation Coefficients (STDC)

In this section, we focus on converging SDC and STC to create the Spectra-Temporal Deviation Coefficients (STDC) fea-

ture set. Given the distinct natures of SDC and STC, we address the range disparity by applying a tailored normalization technique that ensures both sets of coefficients are within a compatible range. The normalized coefficients are then processed through an autoencoder-decoder network, which distils the robust representation of spectra-temporal cues. The reconstruction process of the STDC feature set also aids in alleviating the challenges posed by sparsity in STC features before normalization.

3. EXPERIMENTATION AND RESULTS

3.1. Dataset and Implementation Details

We used several challenging datasets (ASVspoof2019, VSDC, partial spoofs (Utterance-based), ASVspoof2021 and in-the-wild audio deepfakes (IWA)) to evaluate the proposed method [10]. We used the training subset of Asvspoof2019 for training, development subset for validation and testing subset for evaluation of the system. To address the data imbalance in ASVspoof2019 and partial spoof datasets, we applied five augmentation techniques as follows: high-pass filtering, low-pass filtering, compression, time and pitch shift, and reverberation. For our backend classifiers, we used a batch size of 32, the Adam optimizer with an initial learning rate of $1e^{-4}$ and a weight decay of 0.001. Models were trained for 50 epochs using cross-entropy loss.

3.2. Experimental Results

3.2.1. Performance Analysis of the SDC with Different Classifiers

We have evaluated the performance of the proposed SDC features with different machine learning (ML) and residual-based classifiers, and the results are presented in Table 2. It is observed from the results that SDC features performed well with both ML and residual classifiers, with the best performance achieved with Ensemble and SE-ResNext18 classifiers. The lower EERs show the efficiency of the presented coefficients and their potential standalone use for voice spoofing attack detection.

3.2.2. Performance Analysis of STDC with Different Voice Spoofing Datasets

We choose the best-performing back-end classifier (SE-ResNeXt18) from Table 2 and evaluate the performance of the proposed system with different datasets. Results are shown in Table 3, indicating performance improvement when spectral coefficients converge with temporal coefficients. Specifically, EER improves from 0.25 to 0.22, 0.60 to 0.52, 3.70 to 3.50, and so on. These results show the significance of incorporating both spectral and temporal coefficients.

3.2.3. Comparison with Existing SOTA Methods

We evaluate our proposed methods against recent voice spoofing countermeasures, addressing four distinct attack types: LA, PA, and fully and partially deepfake. To our knowledge, this is one of the first comprehensive approach to tackle these four attack categories simultaneously. Moreover, we compared our solution to specific attack-focused methods, such as ASVspoof2019 (LA+PA) in Table 4, ASVspoof2021 in Table 5, partial-spoof in Table 6, and IWA in Table 7. Our method outperforms existing state-of-the-art methods. Though the performance of the method on specific dataset (IWA) and some attacks (PSF) is slightly higher, it exhibits superior generalizability across a wide range of attacks.

Classifiers	ASV-19		ASV-21		PSF	VSDC	IWA
	LA	PA	LA	DF			
Random Forest	0.49	1.20	4.37	5.11	7.90	2.30	29.9
KNN	0.28	1.00	4.17	5.20	6.11	1.89	25.5
SVM	0.22	0.70	4.90	3.95	5.95	1.02	70.7
Logistics Regression	0.30	0.80	3.95	3.90	6.30	2.15	90.0
Naive Bayes	0.31	0.75	4.98	4.50	6.50	2.40	95.5
Decision Tree	0.45	0.90	5.30	5.50	7.11	3.90	19.0
Ensemble	0.26	0.63	3.79	3.40	6.02	2.01	40.0
ResNet18	0.28	0.60	4.01	3.30	5.95	1.56	45.0
SE-ResNet18	0.29	0.63	3.90	3.40	5.98	1.10	35.3
ResNext18	0.25	0.65	3.98	3.35	6.00	1.50	40.4
SE-ResNext18	**0.25**	**0.60**	**3.70**	**34.1**	**5.98**	**0.95**	**32.23**

Table 2: Performance analysis of spectral deviated coefficients with different machine learning and deep learning back-end classifiers.

Table 3: Performance analysis of Spectra-Temporal Deviated Coefficients against different datasets (Lower is better).

Performance	ASV-19		ASV-21		PSF	VSDC	IWA
	LA	PA	LA	DF			
EER	0.22 ± 0.1	0.52± 0.23	3.50 ± 1.25	3.20± 1.30	5.90± 1.50	0.80± 0.15	30.0± 2.50
Accuracy (%)	98.5	98.0	95.5	95.0	93.5	98.5	98.5

Study	Method	ASV-19	
		LA	PA
[11]	CQCC-GMM	9.87	11.04
[11]	LFCC-GMM	11.96	13.54
[12]	FBCC-GMM	6.16	10.36
[13]	SE-Res2Net50	2.86	1.00
[14]	LFCC-CNN	9.09	2.01
[15]	CQT-DCT-LCNN	1.84	0.54
Ours	**STDC+SE-ResNeXt18**	**0.22**	**0.52**

Table 4: Comparison of proposed method with existing methods on ASVspoof2019 dataset (Lower is better).

Study	Method	ASV-21	
		LA	DF
[16]	wav2vec 2.0	1.19	4.38
[17]	LCNN+ResNet+RawNet	1.32	15.64
[18]	ECAPA-TDNN (Ensemble)	5.46	20.33
[19]	ResNet (Ensemble)	3.21	16.05
[20]	W2V2 (fixed)+LCNN+BLSTM	10.97	7.14
[20]	W2V2 (finetuned)+LCNN+BLSTM	7.18	5.44
Ours	**STDC+SE-ResNeXt18**	**3.50**	**3.20**

Table 5: Comparison of proposed method with existing methods on ASVSpoof2021 dataset.

Study	Method	EER
[6]	LCNN	6.19
[6]	SELCNN	6.33
[6]	H-MIL (Ensemble)	5.96
[6]	LS-H-MIL	5.89
[9]	LCNN + LSTM	8.61
[9]	SELCNN(2)+LSTM	7.69
Ours	**STDC+SE-ResNeXt18**	**5.90**

Table 6: Comparison of proposed method with existing models on partial spoof dataset (Lower is better).

Study	Method	EER
[7]	LCNN+LSTM	33.0
[8]	RawGAT	53.00
[8]	RawNet2	51.00
[3]	ECAPA-TDNN	30.3
[4]	H/ASP	27.2
[5]	ClovaAI	36.3
Ours	**STDC+SE-ResNeXt18**	**30.3**

Table 7: Comparison of proposed method with existing models on In-the-Wild Audio deepfake dataset (Lower is better).

4. CONCLUSION AND FUTURE WORK

We have presented a spectra-temporal approach for the detection of a wide range of voice spoofing attacks. Our method incorporates SDC, STC, and STDC obtained through segment-level and utterance-level patterns. Our method successfully addresses various voice spoofing attacks, such as logical, physical, full, and partial deepfake attacks, within a unified framework. The effectiveness of our proposed method has been rigorously evaluated against state-of-the-art unified classifiers, highlighting its potential to enhance voice spoofing detection across a wide range of voice attack scenarios.

5. REFERENCES

[1] Awais Khan, Khalid Mahmood Malik, James Ryan, and Mikul Saravanan, "Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures," *Artificial Intelligence Review*, pp. 1–54, 2023.

[2] Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi, Jose Patino, and Nicholas Evans, "An initial investigation for detecting partially spoofed audio," *arXiv preprint arXiv:2104.02518*, 2021.

[3] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[4] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han, "In defence of metric learning for speaker recognition," *arXiv preprint arXiv:2003.11982*, 2020.

[5] Hee Soo Heo, Bong-Jin Lee, Jaesung Huh, and Joon Son Chung, "Clova baseline system for the voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2009.14153*, 2020.

[6] Yupeng Zhu, Yanxiang Chen, Zuxing Zhao, Xueliang Liu, and Jinlin Guo, "Local self-attention based hybrid multiple instance learning for partial spoof speech detection," *ACM Transactions on Intelligent Systems and Technology*, 2023.

[7] Xin Wang and Junich Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," *arXiv preprint arXiv:2103.11326*, 2021.

[8] Hemlata Tak, Jee-weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," *arXiv preprint arXiv:2107.12710*, 2021.

[9] Lin Zhang, Xin Wang, Erica Cooper, and Junichi Yamagishi, "Multi-task learning in utterance-level and segmental-level spoof detection," *arXiv preprint arXiv:2107.14132*, 2021.

[10] Awais Khan and Khalid Mahmood Malik, "Securing voice biometrics: One-shot learning approach for audio deepfake detection," in *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2023, pp. 1–6.

[11] Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Andreas Nautsch, Xin Wang, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, and Kong-Aik Lee, "Introduction to voice presentation attack detection and recent advances," *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*, pp. 339–385, 2023.

[12] Suvidha Rupesh Kumar and B Bharathi, "A novel approach towards generalization of countermeasure for spoofing attack on asv systems," *Circuits, Systems, and Signal Processing*, vol. 40, pp. 872–889, 2021.

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

[14] Joao Monteiro, Jahangir Alam, and Tiago H Falk, "Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers," *Computer Speech & Language*, vol. 63, pp. 101096, 2020.

[15] Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov, "Stc antispoofing systems for the asvspoof2019 challenge," *arXiv preprint arXiv:1904.05576*, 2019.

[16] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jeeweon Jung, Junichi Yamagishi, and Nicholas Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," *arXiv preprint arXiv:2202.12233*, 2022.

[17] Anton Tomilov, Aleksei Svishchev, Marina Volkova, Artem Chirkovskiy, Alexander Kondratev, and Galina Lavrentyeva, "STC Antispoofing Systems for the ASVspoof2021 Challenge," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 61–67.

[18] Xinhui Chen, You Zhang, Ge Zhu, and Zhiyao Duan, "Ur channel-robust synthetic speech detection system for asvspoof 2021," *arXiv preprint arXiv:2107.12018*, 2021.

[19] Tianxiang Chen, Elie Khoury, Kedar Phatak, and Ganesh Sivaraman, "Pindrop labs' submission to the asvspoof 2021 challenge," *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pp. 89–93, 2021.

[20] Xin Wang and Junichi Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," *arXiv preprint arXiv:2111.07725*, 2021.