

Audio Anti-Spoofing Detection: A Survey

MENGLU LI, YASAMAN AHMADIADLI, and XIAO-PING ZHANG*, Department of Electrical, Computer and Biomedical Engineering, Toronto Metropolitan University, Canada

The availability of smart devices leads to an exponential increase in multimedia content. However, the rapid advancements in deep learning have given rise to sophisticated algorithms capable of manipulating or creating multimedia fake content, known as Deepfake. Audio Deepfakes pose a significant threat by producing highly realistic voices, thus facilitating the spread of misinformation. To address this issue, numerous audio anti-spoofing detection challenges have been organized to foster the development of anti-spoofing countermeasures. This survey paper presents a comprehensive review of every component within the detection pipeline, including algorithm architectures, optimization techniques, application generalizability, evaluation metrics, performance comparisons, available datasets, and open-source availability. For each aspect, we conduct a systematic evaluation of the recent advancements, along with discussions on existing challenges. Additionally, we also explore emerging research topics on audio anti-spoofing, including partial spoofing detection, cross-dataset evaluation, and adversarial attack defence, while proposing some promising research directions for future work. This survey paper not only identifies the current state-of-the-art to establish strong baselines for future experiments but also guides future researchers on a clear path for understanding and enhancing the audio anti-spoofing detection mechanisms.

Additional Key Words and Phrases: Deepfakes, Speech synthesis, Audio anti-spoofing detection, Spoofing countermeasures, ASV

1 INTRODUCTION

Deep learning (DL) techniques have significantly advanced the creation of spoofing speech attacks, commonly referred to as "Deepfake". Deepfake audio holds the potential to propagate misinformation, for example defaming the credibility of prominent figures, leading to political insecurity, fake news, and manipulation of public opinion [37, 284]. Moreover, the rapid growth of Deepfake audio synthesis algorithms also puts voice-enabled devices at risk since the synthesized voices can maliciously take over the control of a device.

The primary techniques of speech synthesis are Text-to-Speech (TTS) and Voice Conversion (VC). TTS models take the given text characters as input and utilize vocoders to generate a natural-sounding speech that follows the linguistic rules of the input text. The dominant TTS algorithms are typically structured as autoregressive-based models, such as WaveNet [214], Tacotron [232] or GAN-based architecture, like HiFi-GAN [107]. On the other hand, VC attacks alter original speech to mimic the voice of a specified target speaker while preserving linguistic information. Notably, the source speech for VC models can originate from a TTS algorithm [255]. As audio spoofing threats continue to emerge, the series of ASVspoo (Automatic Speaker Verification Spoofing and Countermeasures) [142] and ADD (Audio Deep Synthesis Detection) [266] challenges have been developed and played a pivotal role in fostering the development of advanced algorithms to combat audio spoofing attacks. In this survey, we focus on reviewing and analyzing the recent advanced anti-spoofing countermeasures (CMs) targeting TTS and VC attacks across diverse application scenarios, extending beyond the scope of the Automatic Speaker Verification (ASV) systems. Nonetheless, the development of CMs empowers ASV systems against potential threats.

*Corresponding Author

Several surveys on audio anti-spoofing detection have been published in recent years. The main differences between our survey and the existing ones are summarized as follows.

- **Broader definition of spoofed audio.** We adopt a more inclusive definition of spoofed audio, to include both the entire fake audio clips generated by the TTS/VC algorithms and the partial spoofed audio. We evaluate the detection models that specifically target partially fake portions within an audio clip, which has not been addressed in the previous surveys. We also analyze speaker-aware audio anti-spoofing models, exploring the integration of speaker verification and spoofing CMs to protect the ASV systems.
- **Optimization and performance enhancement techniques during the training process.** While the existing surveys primarily concentrate on the architecture of detection algorithms, including feature engineering and classifier structures, our survey extends the analysis to include optimization and training techniques. We highlight the significance of factors like data augmentation and the choice of loss function, which can substantially impact the detection performance. By examining these techniques collectively, we equip researchers with a comprehensive toolkit for developing enhanced algorithms.
- **Explorations of emerging research topics.** The publications in the field of audio Deepfake have increased significantly in recent years, leading to the emergence of advanced topics like adversarial attack defense and cross-dataset evaluation. Our survey provides a thorough review and evaluation of newly published articles, ensuring up-to-date coverage of these topics.
- **Emphasis on open-source availability.** Open source plays a pivotal role as it fosters technological advancement. Our survey places significant emphasis on providing open-source information for all reviewed models and datasets.

More specifically, several existing surveys [95, 103, 160, 283] provide literature reviews on Deepfake media content, with an emphasis on image and video aspects rather than audio. Besides, specific surveys addressing audio anti-spoofing differ from this paper in several aspects. For example, Mcuba et al. [161] mainly focus on reviewing the CNN-based classifiers. Almutairi et al. [6] and Cuccovillo et al. [44] highlight the open challenge in the current solutions rather than providing a detailed comparison of state-of-the-art (SOTA). Wang et al. [228], and Dixit et al. [55] provide an evaluation of the detection methods specifically targeting TTS/VC-generated fully spoofed audio clips. More recently, Khan et al. [102] and Yi et al. [267] briefly touch upon partially fake anti-spoofing detection in their surveys but do not offer a systematic discussion and evaluation of the specific algorithms involved. Furthermore, the existing work also lacks reviews on the optimization techniques that can be applied to the training process for performance enhancement.

Therefore, we provide an in-depth review of various aspects of audio anti-spoofing technology, including the elements of detecting architecture, prevailing training techniques, methodologies embracing diverse application scenarios, and the latest available datasets. For each aspect, we will discuss the detailed design, evaluate the performance, address current limitations and explore future directions. We aim to provide a thorough understanding of the broader picture for preventing malicious audio Deepfakes, serving as a valuable reference guide for future researchers. Specifically, the contribution of our survey can be summarized as follows:

- We present a comprehensive review of each building block of developing audio anti-spoofing detection algorithms and provide the evaluation of performance on both fully and partially spoofing scenarios.
- We are the first to evaluate the effectiveness of optimization techniques applied in the model training process, such as data augmentation, activation functions and loss functions. We

also address the current stage of emerging research tasks, including transferring learning and the explainability of the detection architecture.

- We provide open-source information regarding SOTA and benchmarking datasets, which gives a feasible guide to achieve reproductions.
- We review the existing challenges faced by SOTA models and propose future research directions for advancing audio anti-spoofing detection.

2 DATASETS AND EVALUATION METRICS FOR AUDIO ANTI-SPOOFING

In this section, we present a detailed summary of the datasets and evaluation metrics utilized to address the challenge of audio spoofing.

2.1 Datasets

We provide the details and discuss the characteristics of the most recent widely utilized audio datasets, categorized into three main types: fully spoofed, partially spoofed, and fully real datasets. The fully spoofed datasets typically include both bona fide and spoofed speech generated by TTS and/or VC algorithms. Partially spoofed datasets involve replacing partial segments of the original real utterances with synthesized or manipulated audio. Fully real datasets are crucial for training audio spoofing algorithms or they serve as bona fide samples in the first two categories of datasets. Comprehensive information regarding fully spoofed and partially spoofed datasets is presented in Table 1.

2.1.1 Fully spoofed audio datasets.

ASVspoof2019-LA[255] This dataset is divided into two subsets, logical access (LA) and physical access (PA), with PA featuring the replay-spoof speech and LA containing TTS and VC-generated spoofed speech, all derived from the VCTK database [256]. This survey mainly focuses on the LA subset. The evaluation set in this dataset contains spoofed speech created by 13 unseen algorithms in the training or development set, to evaluate the generalizability of anti-spoofing detection algorithms. It's important to note that all data in the dataset is clean, without noise, or channel variation, which may lead to a detachment from real-world conditions.

ASVspoof2021-LA[142] This dataset is an extension version of the ASVspoof2019-LA, aiming to bridge the gap between ideal experimental and real-world conditions. The evaluation data are across the real telephone systems, incorporating various codecs, transmission channels, bitrates and sample rates. Typically, anti-spoofing algorithms are trained using the training and development set of ASVspoof2019-LA, and then evaluated using ASVspoof2021-LA to test their generalizability towards unknown channel variations. There is no additive noise in this dataset.

ASVspoof2021-DF[142] The Deepfake Speech (DF) subset is newly introduced in ASVspoof2021, distinct from ASV systems. Both bona fide and spoofed speech utterances in the DF track are processed with different lossy codecs, potentially introducing distortion. This DF subset contains the audio clips from the ASVspoof2019-LA evaluation set, along with data from Voice Conversion Challenge (VCC) 2018 [143] and 2020 [268] databases. As a result, the DF evaluation set is generated by more than hundreds of different TTS and VC spoofing attack algorithms under various compression conditions as well as different source domains.

FakeorReal-original (FoR)[183] FoR-original is an English audio dataset that contains both bona fide and spoofed speech generated by diverse TTS algorithms. The FoR dataset provides three publicly available versions, each with different pre-processing methods.

WaveFake[63] WaveFake is a spoofed audio dataset, generated by six different GAN-based TTS algorithms across two languages, English and Japanese. There is no additive noise in this dataset and it only contains two speakers.

Table 1. Statistics of Datasets for Audio Anti-Spoofing Detection

Dataset	Year	Accessibility	Language	Spoofed type	Unseen attacks*	#spoofed methods	Condition	Format	Sample rate	# real	# spoofed	# Male speaker	#Female speaker
Fully spoofed datasets													
ASVspoof 2015	2015	Yes	English	TTS, VC	Yes	10	Clean	Flac	16kHz	16651	246500	45	61
ASVspoof 2019-LA	2019	Yes	English	TTS, VC	Yes	19	Clean	Flac	16kHz	10256	90192	20	28
FoR -original	2019	Yes	English	TTS	No	7	Clean	WAV	Multiple	108256	87285		173
ASVspoof 2021-LA	2021	Yes	English	TTS, VC	Yes	19	Codec	Flac	Multiple	14816	133360	30	37
ASVspoof 2021-DF	2021	Yes	English	TTS, VC	Yes	100+	Codec	Flac	Multiple	14869	519059	43	50
FMFCC-A	2021	Yes	Chinese	TTS, VC	Yes	13	Noisy, Codec	WAV	16kHz	10000	40000	58	73
WaveFake	2021	Yes	English, Japanese	TTS	No	7	Clean	WAV	16kHz	0	117985	0	2
ITW	2022	Yes	English	Not provided			Noisy	WAV	16kHz	19963	11816		58
TIMIT-TTS	2022	Yes	English	TTS	No	12	Noisy, Codec	WAV	16kHz	0	5160		46
ADD2022-LF	2022	Restricted	Chinese	TTS, VC	Yes	Unknown	Noisy	WAV	16kHz	36953	123932	40	40
Latin - American	2022	Yes	Spanish	TTS, VC	No	6	Clean	WAV	48kHz	22816	758000	78	84
CFAD	2023	Yes	Chinese	TTS, VC	Yes	12	Noisy, Codec	WAV	16kHz	38600	77200		1212
MLAAD	2024	Yes	23	TTS	No	54	Clean	WAV	22kHz	0	76000	Not provided	
Partially spoofed datasets													
Partial Spoof**	2021	Yes	English	TTS, VC	Yes	9	Clean	Flac	16kHz	12483	108978	Not provided	
HAD**	2021	Restricted	Chinese	TTS	Yes	Unknown	Clean	WAV	44.1kHz	53612	753612	43	175
Psynd	2022	Restricted	English	TTS	No	1	Codec	WAV	24kHz	30	2371	537	507
ADD2022-PF	2022	Restricted	Chinese	TTS, VC	No	Unknown	Clean	WAV	16kHz	23897	127414	Not provided	
ADD2023-PF**	2023	Restricted	Chinese	TTS, VC	Yes	Unknown	Noisy, Codec	WAV	16kHz	55468	65449	Not provided	

*Unseen attacks refer to spoofed attacks present in the evaluation set but not included in training set.

**These datasets also offer fine-grain labels at segment-level for partial spoofing.

In-the-Wild (ITW)[165] ITW is a dataset of audio Deepfakes with corresponding bona fide audio for English-speaking celebrities and politicians. Both bona fide and spoofed audio samples are collected from publicly available sources such as social networks and video streaming platforms, potentially containing background noises. This dataset is intended to evaluate the generalizability of detection models, including cross-dataset evaluation.

TIMIT-TTS[190] TIMIT-TTS is a synthetic speech dataset containing 12 SOTA TTS algorithms. All selected TTS algorithms are spectrogram generators, which keep the differences between the generated speech primarily attributable to the vocoders. Various post-processing techniques, including adding Gaussian noises, applying MP3 codecs and adding reverberation compression, are applied to reduce the audio quality and hide some artifacts.

FMFCC-A[294] FMFCC-A is a publicly available Mandarin audio spoofed dataset generated by both TTS and VC algorithms. The entire dataset is partitioned into the training, development

and evaluation sets, with the evaluation set featuring speech synthesis algorithms unseen by the training set. Also, half of the evaluation dataset is randomly selected to process a compression-decompression operation or add Gaussian noise to enhance its diversity.

Chinese Fake Audio Detection (CFAD)[150] CFAD is another publicly available Mandarin audio spoofed dataset consisting of twelve different spoofing algorithms including both TTS and VC types. This dataset offers two notable advantages over the FMFCC-A dataset. Firstly, bona fide data in CFAD is sourced from six different domains to prevent bias, whereas FMFCC-A collects from a single source. Secondly, CFAD provides detailed labelling, including information on the spoofed type, real data source, noise type, signal-to-noise ratio, and media codec types.

ADD2022-LF [265] The low-quality (LF) track of the ADD2022 challenge focuses on utterances with various real-world noises and background music effects. This dataset is inaccessible online.

Latin-American Voice Anti-spoofing [210] The dataset utilizes TTS and VC algorithms to generate spoofed speeches with five different accents of Latin-American Spanish.

Multi-Language Audio Anti-spoofing (MLAAD)[166] MLAAD is a newly published spoofed audio dataset created using 54 TTS models across 23 languages. Its novelty is reflected in its variety of languages. This dataset can be utilized either as new out-of-domain test data for existing anti-spoofing models or as an additional training resource.

2.1.2 Partially spoofed audio datasets.

Partial Synthetic Detection (Psynd)[275] The data samples in this dataset are real utterances injected with synthetic speech segments closely resembling the target speakers, generated by multi-speaker TTS algorithms. In training, validation, and preliminary test data, each utterance incorporates one single fake segment. Special cases, such as fully faked, fully real, and multi-fake segments, are stored in the special test set.

PartialSpoof[278] The PartialSpoof dataset contains spoofed speech with varying proportions of spoofed audio segments within a single utterance. This is achieved by pairing speech utterances, with one entirely generated using TTS or VC, and another being an original bona fide speech utterance. Short segments within the pairs are randomly substituted with different lengths. Notably, this dataset offers fine-grain labels, including segmental-level labels at different temporal resolutions.

ADD2022-PF[265] The Partially Fake audio detection (PF) track in the ADD2022 challenge dataset contains fake utterances generated by replacing the partial segments of the original genuine utterances with real as well as synthesized audio. The details of generation algorithms for the spoofed segments are not provided.

ADD2023-PF[266] The PF track is extended in the ADD2023 challenge, which focuses on locating the manipulated regions in partially fake audio in addition to spoofing detection. Additive noise and format conversions are also applied to the utterances.

Half-Truth (HAD)[264] The HAD dataset features partially fake speech where a few words in an utterance are altered using TTS generation techniques. This dataset is designed to evaluate anti-spoofing methods and localize partially fake audio. The replaced keywords include entities, such as person, location, organization, and time.

2.1.3 Fully real audio datasets.

Voice cloning toolkit (VCTK)[256] This dataset includes speech data uttered by 110 English speakers with different accents. All speech data is recorded using the same recording setup with 16 bit/s and 48kHz.

LibriSpeech[175] This dataset is a corpus of approximately 1000 hours of reading English speech with a sampling rate of 16 kHz.

VoxCeleb2[41] This dataset is a multi-language dataset that contains over one million human voices for 6,112 celebrities, extracted from videos uploaded to YouTube. The speech data are captured with real-world noise including laughter, cross-talk, channel effects, music, and more.

LJ Speech[90] This dataset consists of 13,100 short audio clips of a single speaker reading passages from 7 non-fiction English books.

AISHELL-3[192] This dataset is a high-fidelity multi-speaker Mandarin speech corpus. The corpus contains roughly 85 hours of emotion-neutral recordings spoken by 218 native Chinese Mandarin speakers and a total of 88035 utterances.

2.2 Evaluation metrics

We examine the evaluation metrics utilized in the audio anti-spoofing literature, discussing their strengths and limitations. Furthermore, we emphasize metrics tailored to address the specific challenges of detecting partially spoofed content.

Equal Error Rate (EER) The EER is one of most widely used evaluation metrics for audio spoofing CMs. It represents the CM threshold where the false acceptance rate equals the false rejection rate. A lower EER value signifies better performance. This metric offers a more comprehensive and objective assessment compared to accuracy, particularly in scenarios with unbalanced evaluation datasets. Notably, EER serves as the evaluation metric in the ASVspoof and ADD challenge series.

F1-score The F1-score is another metric commonly used in binary classification problems to handle unbalanced evaluation datasets. It represents the harmonic mean of precision and recall, effectively considering both the false acceptance rate and the false rejection rate.

Accuracy Accuracy is the most intuitive metric to reflect the detection performance. However, it can be biased by unbalanced conditions within the evaluation dataset.

Tandem Detection Cost Function (t-DCF) [106] The t-DCF metric is developed during the ASVspoof2019 challenge as an ASV-centric evaluation method. It shifts the focus from the spoofing CMs alone to offer a more comprehensive assessment for ASV attack detection. Recognizing that spoofing CMs and ASV systems operate under different hypotheses and objectives, the t-DCF considers both systems combined in cascaded order. It reflects the cost of detection decisions made by the combination of ASV and CM in a Bayesian sense.

Range-based EER [279] This metric is specifically proposed to evaluate segment-level spoofing detection for the partially spoofed task. Unlike traditional EER measurements that compare discrete predicted segment scores and their corresponding segment-level labels, which may be influenced by the resolution of utterances, a range-based EER can be utilized. This approach measures the duration of misclassified regions between references and hypotheses of each trial with a finer resolution.

3 FULLY SPOOFED DETECTION

In this section, we comprehensively evaluate every component within the detection pipeline for fully spoofed audio, including algorithm architectures and training optimization techniques. Audio anti-spoofing detection models are typically structured by two modules: front-end feature extraction and back-end classifier. However, End-to-end (E2E) architectures have gained more attention due to their capability to avoid information loss caused by pre-defined feature extraction. We evaluate recent advancements in feature engineering, classifier development, and E2E architectures. Furthermore, we assess various training optimization techniques mentioned in the literature, including data augmentation, loss functions, and activation functions. Our focus lies on elucidating the effectiveness of each optimization technique to enhance the performance of spoofing CMs.

Table 2. The Performance of Single-System State-of-the-art Models on the Series of ASVspoof Evaluation Sets

Publication	Data augmentation	Feature	Classifier	Loss function	# Params	ASVspoof			Accessibility
						19-LA	21-LA	21-DF	
[291]	INTERSPEECH'21	w/o	Mel-Spec on 0-4kHz	SE-ResNet-18	AM-Softmax	1.1M	1.14	-	No
[204]	INTERSPEECH'21	channel masking	RawNet2*	GAT	CE	440K	1.06	6.92	Yes ¹
[71]	INTERSPEECH'21	channel masking	SincNet	Raw PC-DARTS	MSE	24.4M	1.77	6.43	Yes ²
[87]	SPL'21	mix-up	E2E: CNN→ResNet→MLP	CE	350M	1.64	-	-	Yes ³
[65]	ICASSP'22	w/o	FastAudio	ECAPA-TDNN	CE	Unknown	1.54	-	Yes ⁴
[119]	DSP'22	w/o	L-VQT	DenseNet	CE	338K	2.19	-	No
[212]	ICPR'22	w/o	RawNet2+(CQT→ECAPA-TDNN)	CNN→MLP	CE	7.19M	1.11	-	Yes ⁵
[114]	INTERSPEECH'22	w/o	wav2vec2.0-XLSR	MLP	CE	317M	0.31	-	No
[59]	INTERSPEECH'22	w/o	wav2vec2.0-960	MLP	CE	Unknown	0.40	-	No
[39]	INTERSPEECH'22	frequency masking	CQT-Spec	LCNN	CE	135K	1.35	-	No
[227]	ODYSSEY'22	w/o	wav2vec2.0-XLSR	Bi-LSTM →MLP	CE	317M	1.28	6.53	No
[209]	ODYSSEY'22	RawBoost	wav2vec2.0-XLSR	AASIST	CE	Unknown	-	0.82	2.85 Yes ⁶
[144]	DDAM'22	RawBoost	ImageNet + Jitter + Shimmer	MLP	AM-Softmax	Unknown	0.87	10.06	27.08 No
[217]	DDAM'22	w/o	wav2vec2.0-Large	DARTS	Unknown	Unknown	1.08	-	7.89 No
[92]	ICASSP'22	w/o	RawNet2	GAT	CE	297K	0.83	5.59	- Yes ⁷
[117]	SPL'22	w/o	LFCC	OCT	Focal loss	250K	1.06	-	- No
[132]	APSIPA'22	adding noise, RIRs	wav2vec2.0	LCNN	CE	Unknown	0.24	-	- No
[100]	MAD'23	w/o	Mel-spec + Spec-Env + Spec-Contrast	Transformer →CNN	CE	603K	0.95	-	- No
[218]	INTERSPEECH'23	w/o	Duration + pronunciation + wav2vec2.0-XLSR	LCNN →Bi-LSTM →MLP	CE	Unknown	1.58	-	- No
[151]	SPL'23	w/o	(LFCC →ResNet) + (CQT-Spec →ResNet)	GRL →MLP	CE	Unknown	0.80	-	- Yes ⁸
[139]	ICASSP'23	w/o	RawNet2	Rawformer	CE	370K	0.59	4.98	4.53 Yes ⁹
[158]	ICASSP'23	FIR filter	wav2vec2.0-XLSR	MLP	OC-Softmax	300M	-	3.54	6.18 No
[32]	ICASSP'23	time & frequency masking	LFB-Spec	GCN	CE	Unknown	0.58	-	- No
[276]	ALGORITHM'23	RawRoost	wav2vec 2.0	Transformer	CE	Unknown	-	1.18	4.72 No
[101]	ICASSP'24	FIR filter, codec, noises, shift	SDC + Bi-LSTM	Auto-encoder →SE-ResNeXT	CE	Unknown	0.22	3.50	3.41 No

¹ <https://github.com/eurecom-asp/RawGAT-ST-antispoofing>² <https://github.com/eurecom-asp/pc-darts-anti-spoofing>³ <https://github.com/gghua-ac/end-to-end-synthetic-speech-detection>⁴ <https://github.com/magnumresearchgroup/Fastaudio>⁵ <https://github.com/magnumresearchgroup/AuxiliaryRawNet>⁶ https://github.com/TakHemlata/SSL_Anti-spoofing⁷ <https://github.com/clovaai/aasist>⁸ <https://github.com/imagecbj/End-to-End-Dual-Branch-Network-Towards-Synthetic-Speech-Detection>⁹ <https://github.com/rst0070/Rawformer-implementation-anti-spoofing>

* RawNet2 consists of a learnable SincNet filter and six ResNet blocks

The evaluation metric is EER (%). "-" indicates that the authors do not report the performance with the corresponding dataset. The bold values refer to the best performance on the same dataset. "+" indicates multiple techniques processed in parallel, while "→" denotes sequential order.

3.1 Feature Engineering

We categorize the current methodologies of feature extraction into three groups: hand-crafted traditional spectral features, deep-learning features, and other analysis-oriented approaches, as summarized in TABLE 3.

3.1.1 Hand-crafted spectral features. Hand-crafted features have been demonstrated as a strong baseline for audio anti-spoofing detection, providing a reliable foundation for capturing discriminative patterns of artifacts.

Magnitude-based spectral coefficient The literature shows that the majority of front-end features are derived from the magnitude/power spectrum, where the power spectrum is the square of the magnitude spectrum. Short-term magnitude spectral features are commonly obtained from discrete Fourier transforms (DFT), such as Mel frequency cepstral coefficient (MFCC) [50], inverted Mel frequency cepstral coefficient (IMFCC) [28], linear frequency cepstral coefficients (LFCC) [4], and rectangular filter cepstral coefficients (RFCCs) [85], and linear prediction cepstral coefficients (LPCC) [66]. Sahidullah et al. [188] are the first to compare nine short-term magnitude-based spectral features alongside their first- and second-order derivatives, highlighting the benefits of incorporating dynamic features in capturing temporal changes. Recent advancements have focused on enhancing conventional coefficients by refining windowing techniques and filter configurations during the extraction process. Mewada et al. [163] propose to utilize a Gaussian filter to obtain the IMFCC because of their ability to capture more global information compared to linear or triangular filter banks. Liu et al. [137] extract and stack short-time Fourier transform (STFT) feature maps with varying time-frequency resolutions by using different window lengths and frame shifts. Additionally, [137] notes that the longer window length performs more effectively than the shorter one. Gammatone cepstral coefficient (GTCC) [26] utilizes an Equal Rectangular Bandwidth (ERB) frequency scale, which is more robust to the noise compared to the traditional Mel-scale. Gasenzer et al. [69] propose the wavelet packet transform (WPT) as an alternative to the DFT since the WPT provides a high resolution in the high-frequency part. The experiments suggest that the WPT outperforms STFT in the situation of low sampling rates or compression.

In contrast to short-term spectral features derived from fixed short window lengths, long-term window transforms have been proposed to capture long-range information and achieve higher frequency resolution. Constant-Q cepstral coefficient (CQCC) [207] has shown its effectiveness in audio anti-spoofing detection by providing higher frequency resolutions at lower frequencies and higher temporal resolution at higher frequencies. Variants of CQCC are continuously proposed. Li et al. [127] modify the CQCC by incorporating a block transform, segmenting the CQT log power spectrum into overlap blocks and performing the discrete cosine transform (DCT) individually on each block. It leads to a 36% improvement over conventional CQCC on the ASVspoof2015 dataset. Yang et al. [259] propose to keep the information from the octave power spectrum in addition to CQCC from the linear power spectrum. The CQT-based power spectrum is inverted in [258] to emphasize the high-frequency information. Kwak et al. [110] explore the impact of frequency variations by experimenting with different values for minimum central frequency and total numbers of frequency bins while fixing the number of bins per octave and hop size. By setting the minimum central frequency to 1Hz and the total number of frequency bins to 100, it receives an EER of 2.19% on the ASVspoof2019-LA set. In addition to CQT, Li et al. [119] introduce a long-term variable Q transform (L-VQT), where the frequencies vary as a power function rather than exponential as in CQT, aiming to capture better high-frequency information and detect artifacts created by commonly-used vocoders like WaveNet, even in noisy conditions. Apart from CQT-based features, Gao et al. [67] utilize global 2D-DCT on Mel-scale magnitude spectrum across both temporal and frequency dimensions to capture long-term modulation artifacts created by frame-level audio generation algorithms.

Phase-based spectral coefficient The spoofed speech often lacks natural phase information, as the human auditory system tends to be less sensitive to phase spectrum characteristics compared to magnitude spectrum features. Therefore, investigating the phase information can be effective in capturing these artifacts in spoofed speech. In addition to short-term phase-based features, such as modified group delay cepstral coefficients (MGDCC) [245], Wang et al. [221] propose a relative phase extraction method aimed to reduce the phase variation, where the peaks of the utterance waveform serve as the center of each window section. Furthermore, Gupta et al. [83]

Table 3. The Categorization of Feature Extraction Methods

Category		Description	Methods
Hand-crafted spectral features	Magnitude/power-based spectral coefficients	Short-term: Computing a frequency domain transform on each temporal window of the audio signal to enhance time resolution at lower frequencies.	MFCC [50], IMFCC [28], LFCC [4], RFCC [85], LPCC [66], SSFC [188], SCFC [188], SCMC [188], Gaussian-IMFCC [163], Multi-resolution STFT [137], GTCC [26], WPT [69], STFT [126], PLP [124], DTW [80], Spec-Env [100], Spec-Contrast [100]
		Long-term: Longer temporal window.	CQCC [207], CQBC [127], eCQCC [259], CLBC [258], L-VQT [119], SCC [200], CQMOC [261], CFCC [177], Global M [67]
	Phase-based spectral coefficients	Working effectively as a complement to the magnitude features. However, it may not be helpful for unknown attacks, especially for VC attacks.	MGDCC [245], APGDF [174], Relative phase [221], Quadrature phase [83], MMPS [261], IF [177]
	Bispectrum	Describing the higher-order spectral correlation in the Fourier domain; is useful for the known attacks.	Statistics of bispectral correlations [3]
	Image-like	Propagating in time to show the variations in frequencies and intensities of an audio signal in a 2D feature, which are interpreted as images.	STFT-spec [16], Mel-spec [182], CQT-spec [1], E-Spect [246], C-CQT spectrogram [170], LBP [77], MLTP [89], SDC [101]
DL features	Filter-learning features	Utilizing DL techniques to construct learnable filterbanks or approximate the standard filtering process.	nnAudio [36], DNN-FBCC [270], FastAudio [65], SincNet [273], TD-FBanks [272], LEAF [271]
	Supervised embeddings	Constructing deep embeddings using DL models through supervised training	CNN [244], ResNet [193], X-vector [29], auto-encoder [15], Bi-LSTM [101], U-net [30]
	Pre-trained embeddings	Utilizing SSL models or other DL models pre-trained by external large datasets to extract latent representations of the raw audio waveform.	wav2vec2.0 [227], WavLM [298], HuBERT [123], TDNN [154], HiFi-GAN [56], and ImageNet [144]
Analysis-oriented features	Prosody/semantic features	Focusing on the prosody and emotion of the speech sounds, which works effectively on TTS-based spoofed audio, not the VC.	Vocal tract estimation [20], shimmer [120], phoneme duration [218], pronunciation [218], prosody [10], emotion [43], VOT [53], coarticulation [53]
	The impact of silence	Contributing effectively to the current anti-spoofing detection models.	Silence portion [169], BTS-Encoder [57]
	Frequency sub-band feature	Focusing on one or more specific portions of the frequency band, rather than the entire frequency range.	F0 [60], 0-4kHz [291], 4-8kHz [168]
	Other possible directions	Including recent attempts on the development of anti-spoofing features.	Varied input length [226], energy loss [52], face embedding [252], dual channel stereo feature [135], Compressed coding metadata [254]

demonstrate the significance of a quadrature phase over other phase angles by performing Mutual Information-based analysis.

Nevertheless, experiments indicate that relying solely on the phase information may lack discriminative power compared to the magnitude-based information. Therefore, phase information always serves as a complement to magnitude information in anti-spoofing detection. Yang et al. [261] propose a modified magnitude-phase spectrum (MMPS) to collectively capture both magnitude and phase information while preserving the sign of the magnitude part. Kim et al. [105] apply a convolution layer with batch normalization (BN) and Rectified Linear Unit (ReLU) activation to the phase feature to mitigate the high randomness of the phase spectrum before concatenating it with the magnitude feature. Patil et al. [177] integrate phase information by estimating the Instantaneous Frequency (IF) with magnitude information represented by cochlear filter cepstral coefficients (CFCC). They observe that the dynamic variations in the IFs of real speech are substantially larger than those of spoofed speech.

Bispectrum The magnitude/power spectrum lacks sensitivity to higher-order spectral correlations, which is revealed by bispectral analysis. AlBadawy et al. [3] conduct a qualitative assessment

to reveal the difference in both magnitude and phase of bispectrum between real and synthesized speech. Leveraging statistics of the first four moments of bispectral correlations as an eight-dimensional hand-crafted feature achieved high accuracy detection performance with a basic linear classifier, particularly under a high Signal-to-Noise Ratio (SNR) condition. The bispectral features from [3] are combined with power spectrum features, such as STFT and MFCC, which improves the detection performance for the known attacks [21, 196]

Image-like features In audio anti-spoofing detection, magnitude-based spectral coefficients are typically integrated with the magnitude of the audio signal over time to form a spectrogram, a two-dimensional (2D) feature. The spectrogram includes information regarding frequencies and intensities of the audio signal as it propagates in time. Front-end features, such as Mel-spectrogram (Mel-Spec), and CQT-spectrogram (CQT-Spec) are always treated as images and passed to CNN-based back-end classifiers [7, 16, 30, 76, 182]. The spectrogram requires less computation power than extracting the spectral coefficients through DCT while promising detection accuracy [1]. Xiang et al. [246] propose an Efficient Spectrogram (E-Spec), which applies STFT directly to the decoded audio signal along the frequency axis after the compression filterbank. E-Spec approximates the spectrogram of the MP3 speech without decompressing the signal, and it outperforms the original Spectrogram by 11% on the compressed ASVspoof2019-LA evaluation set. Phase information is also integrated with the magnitude spectrogram [93, 179]. Muller et al. [170] introduce a complex-valued CQT (C-CQT) log-spectrogram to embed phase information, which requests modifications to classifier networks, activation functions, and Batch Normalization (BN) to handle complex-valued input and weights. Furthermore, some research applies texture analysis tools to spectrogram-based features, including local binary pattern (LBP) [51, 77], and Modified local ternary patterns (MLTP) [89], along with edge detection tools like Canny [162]. Khan et al. [101] propose a method to slice the log-mel spectrogram into square segments. For each segment, the local deviated pattern (LDP) operation is applied to identify the local higher and lower frequency spectrum, forming local spectral deviation coefficients (SDC) for detecting the frame-level inconsistencies.

3.1.2 Deep-learning (DL) features. With the advancement of deep learning approaches, DL-based structures have been adopted to extract learnable embeddings to describe the underlying characteristics of raw audio, alongside traditional hand-crafted features. Various types of the recent DL-based features are discussed below.

Filter-learning feature DL techniques are involved in approximating the standard filtering process for both STFT-based and First-order Scattering Transform (FST)-based front-ends. Learnable STFT-based features like nnAudio [36], and DNN-FBCC [270] are implemented without constraining the shape of the filter, leading to a larger number of parameters in training and potential overfitting. Fu et al. [65] enhance nnAudio by restricting the filter shape to triangular and making only the filterbanks learnable, resulting in improved performance. On the other hand, learnable FST-based front-ends, such as SincNet [273], utilize a convolutional layer to parameterize the sinc function, effectively acting as a customized filter bank. However, it also may suffer from overfitting by learning the low and high cut-off frequencies during training. In the work of RawNet2 [208], SincNet is utilized to extract the front-end feature directly from the raw audio while fixing the cut-off frequencies to reduce overfitting. The success of RawNet2 makes it one of the most well-known and reproducible models in audio anti-spoofing detection, serving as an official baseline in the ASVspoof challenge series. Furthermore, [204] adds one additional channel dimension to the output of the SincNet front-end to form a time-frequency representation.

Supervised embedding In this category, DL models, such as Deep Neural Networks (DNN), Residual Networks (ResNet), and Recurrent Neural Networks (RNN) are applied directly on the raw audio data or after the hand-crafted feature to construct deep embeddings through supervised

training [15, 48, 101, 193]. Teng et al. [212] choose to use the whole structure of RawNet2 as an encoder for raw waveforms, combined with ECAPA-TDNN [34] as the encoder for the CQT feature. Two embeddings from both encoders are concatenated and encoded by a convolutional layer with BN. Wu et al. [244] construct a stack of convolutional layers and convolutional transpose layers acting as a genuinization transformer, which is similar to an autoencoder to learn the characteristics of bonafide speech and amplify differences between fake and real speech. Most recently, [30] employs a U-net network with attention mechanisms and skip connections, termed Twice Attention U-net (TA-Unet) [186], to process CQT-spectrograms, aiming to prevent overfitting while emphasizing artifact locations on the spectrogram feature. Wang et al. [224] explore ResNet-based DL structures to capture raw layer-wise neuron behaviours. The activated neurons that have better capability to identify the difference between real and fake are passed to the back-end classifier. Performing detection by observing the neuron behaviours is found to be robust to the real-world noise. Alam et al. [2] intergrate higher order statistics (HOS) into the output embedding of the Time delay neural network (TDNN), incorporating statistical information such as skewness and kurtosis to enhance detection performance, beyond mean and standard deviation derived from the statistics pooling layer in TDNN.

Pre-trained embedding Self-supervised learning (SSL) models have demonstrated their capability to generate latent representations of raw audio waveforms. These SSL-based features outperform the hand-crafted acoustic features or other learnable features in various tasks, including speech and emotion recognition. In audio anti-spoofing tasks, research works show that replacing hand-crafted features or SincNet front-ends with pre-trained wav2vec 2.0 features significantly boosts detection performance when employing the same back-end classifier and data augmentation techniques [115, 148, 209]. Furthermore, fine-tuning SSL features along with the classifier during training accelerates convergence and enhances detection performance for both known and unknown attacks [227]. The commonly-used pre-trained SSL models in audio anti-spoofing detection are wav2vec 2.0 [59, 227], WavLM [235, 298], HuBERT [123], and Whisper encoder [99]. Specifically, experiments highlight that pre-trained models trained on diverse speech data sources, such as wav2vec 2.0-Large2 [14] and wav2vec2.0-XLSR [12], achieve better results on out-of-domain samples [227]. Integration of attention mechanisms into SSL-based models further enhances their effectiveness. [158] applies a temporal normalization on the hidden state of each transformer layer in the wav2vec 2.0 model, where each normalized representation is multiplied with a trainable weight during fine-tuning. Zhu et al. [298] assign different weights to each channel of the deep embedding to maximize the effectiveness of specific channels for discriminating spoofed detection.

In addition to SSL-based features, certain other deep-learning architectures pre-trained using external datasets have been reported in the literature as front-end representation extractors, including TDNN [154], HiFi-GAN [56], and ImageNet [144]. However, these pre-trained models are not as effective as SSL models, like wav2vec 2.0, especially for more robust datasets with varied codec conditions.

3.1.3 Other analysis-oriented features. The majority of efforts on feature engineering for audio anti-spoofing detection concentrate on extracting hand-crafted spectral representations or high-level embeddings using DL techniques. At the same time, various other specific directions for feature development have been explored to improve the robustness of anti-spoofing systems, such as analyzing the impact of silence and sub-band frequencies.

Prosody and semantic features Blue et al. [20] suggest that audio Deepfake models may produce significant inconsistencies compared to a regular human vocal tract, such as unnatural vocal tract diameters. Therefore, they develop a mathematical model to estimate the cross-sectional area of the vocal tract at various points along the speaker's airway and design an anti-spoofing detector

capable of detecting TTS speech. Dharmyal et al. [53] investigates microfeatures, including Voicing Onset Time (VOT) and coarticulation, which are associated with voice production mechanisms in humans. The continuous shimmer is proven to reflect the stability of amplitude and frequency perturbation in the voice and is utilized as a feature to distinguish spoofed audios [120, 121].

DL-based prosody and semantic features have also been proposed. Wang et al. [218] incorporate two types of prosodic features with wav2vec 2.0 embeddings: phoneme duration feature and pronunciation features. They are extracted using a pre-trained HuBERT and a Conformer model [81] respectively. Attorresi et al. [10] trains the same structure of the prosody encoder [197] used by Tacotron to enhance prosody in TTS-synthesized speech as the prosodic embedding extractor in the detection process. Conti et al. [43] is the first to use a pre-trained Speech Emotion Recognition (SER) system to extract emotion embeddings based on the semantics for anti-spoofing detection. Like other mentioned prosody and semantic features, this emotion embedding is effective for TTS-generated fake speeches only, rather than VC-based spoofed speech.

The impact of silence Some researchers argue that the silence portion works as a significant feature for the current anti-spoofing detectors [157, 169, 172]. Specifically, the duration proportion of silence plays a significant role in detecting TTS spoofing, while the content of silence is an important factor in detecting VC attacks [288]. It is because TTS algorithms lack the ability to model diverse and accurate pauses, whereas silence portions in VC spoof audios have signal discontinuity compared to bonafide speech. Utilizing voice activity detection (VAD) to detect and remove silent portions leads to performance degradation. To utilize the effectiveness of the silent part, Doan et al. [57] encode the correlation between breathing, talking and silence sounds in audio clips as the front-end feature. It is important to note that all existing research on the impact of silence is conducted using clean data. These findings may not necessarily hold in noisy conditions or when using various codecs.

Frequency sub-band feature Instead of utilizing feature maps covering the entire frequency range, studies have investigated sub-band spectral information to identify specific ranges containing more discriminative information relevant to detecting spoofing speech [199, 201]. Research shows that sub-band features, particularly within the low-frequency band of 0-4kHz, outperform the full-band features against channel effects, codecs and noisy conditions, while the high-frequency part of spectrograms may lead to overfitting [126, 233, 260]. This finding holds significance for the development of resource-constrained anti-spoofing detectors. [60] and [250] further narrow down the low-frequency band to 0-400Hz, focusing on the fundamental (F0) frequency. By only including the F0 sub-band of the log-power spectrogram as the feature, it can still achieve a satisfactory detection result on the ASVspoof2019-LA set with an EER of 1.15%. [168] point out that, in voiced segments, most spectral differences lie within the 0-4kHz frequency band. Conversely, for silence and unvoiced segments, the spectral discriminating features predominantly reside in the 4-8kHz range. This observation may explain performance degradation after silencing removal, particularly when the feature emphasizes high frequencies.

Varied input length To pass inputs into DL architectures in batch, audio inputs are often set to be fixed-size through trimming or padding. However, this approach may lead to information loss or the propagation of irrelevant information [35]. Wang et al. [226] propose to add a pooling layer before DL models to handle varied input lengths, which outperforms fixed-length inputs with the same back-end classifiers. Consequently, more research has started to accept variable-length speech as an input [117, 139, 165].

Other possible directions Other research endeavours have aimed to develop alternative types of features contributing to anti-spoofing detection. Deng et al. [52] investigate the energy loss in pauses between words and the high-frequency range caused by spoofing algorithms. Yadav et al. [254] explore the utilization of compression coding metadata information solely from the compressed

bit-stream as a feature, which overcomes detection problems for compressed speech, such as Advanced Audio Coding (AAC) compressed audio. Xue et al. [252] construct face embeddings from spectrograms to describe speaker information such as gender, and mouth shape, and concatenate face features and audio features as the front-end. Liu et al. [135] convert mono audio signals to dual channels, encoding and processing each channel signal separately to capture different detail cues for anti-spoofing detection.

3.1.4 Performance discussion on feature selection. The current trend in feature engineering for audio anti-spoofing is shifting from hand-crafted features towards deep embedding representations, particularly derived from pre-trained SSL-based models. This transition is motivated by that conventional acoustic features, based on mathematical principles, may not fully capture hidden information from unknown attacks as effectively as learnable features. Learnable features excel in extracting high-level representations from raw audio data, leading to improved performance in cross-dataset testing scenarios. Nevertheless, despite their limitations, hand-crafted features should not be discarded. Hand-crafted features demand fewer computational resources while offering a high degree of interpretability. In contrast, learnable features usually require a longer training time with a larger amount of model parameters. In cases of limited training data quantity or quality, hand-crafted features and SSL-based features employing transfer learning often outperform learnable features. Therefore, a promising direction involves integrating both hand-crafted and learnable features to construct a robust system. Additionally, while utilizing prosodic-based and phase-based features alone may not yield competitive detection outcomes, they offer value as complementary features alongside others, such as magnitude-based spectral features and learnable features.

In the literature, various techniques are employed to enhance performance, such as pre-emphasis. Pre-emphasis involves applying a first-order high-pass filter directly to the speech signal to amplify its high-frequency content, thereby boosting the energy of the sound [26, 88, 100, 101]. This practice is commonly applied directly to speech signals before feature extraction to address issues related to high-frequency noise induced by transmission. Normalization on spectrograms is another widely used operation in the literature, which reflects mainly the tonal characteristics of speakers [30, 76, 100].

3.2 Classifier Architecture

In addition to traditional machine learning classifiers, SOTA anti-spoofing algorithms focus on utilizing DL architectures such as CNN and ResNet as classifiers. We assess the strengths and limitations of different classifier structures, as summarized in TABLE 4. Certain models incorporate multiple DL architectures. For instance, RawNet2 [208] integrates a gated recurrent unit (GRU) layer after ResNet blocks. Here, we categorize these models based on their primary structure.

3.2.1 Traditional Machine Learning (ML) classifiers. Classic ML-based classifiers are commonly used in the early years of audio anti-spoofing detection, including support vector machines (SVM), Gaussian mixture models (GMM), and random forest (RF) [29, 91, 269]. In particular, GMM-based classifiers serve as a fundamental baseline method in the ASVSpoo challenge series.

3.2.2 Convolutional Neural Network (CNN). CNN architecture is well-known for its effectiveness in capturing local and hierarchical features. Lavrentyeva et al. [113] apply the CNN architecture to address the anti-spoofing problem while reducing model size by implementing a Light-CNN (LCNN). LCNN mainly replaces ReLU with Max-Feature-Map (MFM) activation, which selects the maximum value of each of the two feature channels as output, effectively halving the LCNN architecture. The MFM layer also performs feature selection. The effectiveness of LCNN is also

Table 4. The Categorization of Classifiers

Category	Advantages	Disadvantages	Methods
Traditional ML	Light-weight; facilitating easier interpretation of the distribution outcomes	Poor generalization performance on unseen attacks	GMM [269], RF [91], SVM [29]
CNN	Light-weight; Producing promising detection performance	Causing information loss in the frequency domain due to the translation invariant property	LCNN [113], Non-OFD [39], CapsuleNet [147]
ResNet	Enabling architectural adjustments for modifying receptive fields; enhancing generalizability to unseen attacks; accommodating deeper networks	High computational cost; The performance can be highly varied by feature selection	ResNet [7], SE-Net [112], ResMax [110], ResNext [296], Res2Net [128], DenseNet [234], xResNet [25]
GNN	Aggregating all node features for message passing; enhancing the formulation of inter-relationships among frame-level features	Challenging to construct a deep network; high time and space complexity	RawGAT [204], AASIST [92], GCN [32]
Transformer	Effectively capturing long-term dependencies	Potential for overfitting; high computational costs	CCT [18], OCT [117], TFT [235], Rawformer [139]
TDNN	Lightweight; allowing varying input lengths	Unsatisfactory detection performance	ECAPA-TDNN [34], AF-TDNN [243]
DART	Enabling architecture optimization during back-propagation	Performance may be influenced by pre-defined hyperparameters	PC-PARTS [70], Raw PC-PARTS [71], light-DARTS [217]

proven in various pieces of literature [153, 215, 218]. However, the translation invariance property in CNN may lead to information loss in the frequency domain, particularly because different sub-band frequencies contain diverse information. Choi et al. [39] suggest splitting spectrogram inputs along the frequency axis and processing the high-, mid- and low-frequency band by the LCNN separately. Ranjan et al. [181] consider both frequency and temporal information separately by performing CNN modules on these two domains in parallel. Luo et al. [147] modify the dynamic routing strategies in the capsule network to be suitable for audio anti-spoofing detection, emphasizing hierarchical structures of features and spatial information of the artifacts. This approach achieves satisfactory outputs without data augmentation.

3.2.3 Residual Network (ResNet). ResNet is one of the significant variants of CNN architecture, addressing the vanishing gradient problem in a deep network by incorporating skip connections. ResNets are also widely used in audio anti-spoofing tasks and achieve promising outcomes [7, 208]. Recent works have focused on modifying and enhancing the fundamental structure of ResNet. Lai et al. [112] integrate squeeze-and-excitation (SE) blocks with ResNet, forming SE-Nets, to perform dynamic channel-wise feature recalibration. [110] and [115] adapt the MFM activation layer to each ResNet block. termed ResMax. Instead of sequentially connecting all residual-connected convolution blocks, Li et al. [116] propose Deep layer aggregation (DLA) to group them into a tree-like structure, to enhance information fusion across multiple resolutions and integrate local and global information. ResNeXt modifies the ResNet by stacking more subpaths in parallel within each block to learn more diverse features [296]. To improve the information flow in ResNet, DenseNet [40, 45, 234] is proposed to skip connections linking each layer to all layers within the same dense block. DenseNets also have fewer parameters compared to the conventional ResNets.

In addition to the previously discussed techniques, Res2Net stands out as another significant variant of the ResNet architecture [128, 129, 220]. Res2Net modifies the bottleneck block to incorporate a hierarchical residual-like connection. Instead of passing the entire feature map to the convolutional layer as a whole, Res2Net divides the input feature map along the channel dimension into several feature segments of equal size. Before conducting the convolutional operation on each subsegment, the convolutional result from the previous subsegment is added, creating a multi-scale feature

representation. Furthermore, adjustments are made to the addition operation of feature segments in the basic Res2Net structure. For instance, a dynamic modulated (DM) mechanism assigns different weights to both the current and previous feature segments before addition [262]. Dong et al. [58] implement additional convolutional operations with various kernel sizes on the feature segments after addition to gain multi-perspective information (MPIF) from different receptive fields.

3.2.4 Graph Neural Network (GNN). To apply GNN in audio anti-spoofing, the frequency bins and time frames can be utilized as nodes to form a fully connected graph. One common variant of GNN is the graph convolutional network (GCN). Chen et al. [32] divide the spectrogram into grid patches and extracts each patch embedding using a CNN, emphasizing the network's focus on the relationship between patches. Position embeddings are also added to retain positional information. Subsequently, each patch embedding passes through several layers of GCN to aggregate node features within the same time frames or frequency bands.

Graph attention network (GAT) introduces the attention mechanism during node feature aggregation. Tak et al. [205] conduct experiments utilizing several GATs with the conventional handcrafted feature, log-linear filterbank feature (LFB). Then, [204] replace the LFB features with the learnable SincNet front-end, named RawGAT. The GAT-based classifier consists of three components. Initially, the first two GATs individually model the relationships within spectral and temporal domains. Then, these two sub-graphs are fused to facilitate the processing of the third GAT, thereby leveraging complementary information. AASIST [32] is introduced as an enhancement to RawGAT by incorporating a heterogeneity-aware technique to integrate spectral and temporal sub-graphs. In AASIST, each node aggregates information from all other spectral nodes and temporal nodes in the graph, whereas nodes in RawGAT only aggregate information within the spectral and temporal sub-graphs individually. Huang et al. [88] make two adjustments to RawGAT by adding a pre-emphasis module before the SincNet filter to enhance the high-frequency components and replacing BatchNorm with LayerNorm to reduce the impact caused by uneven samples. These two adjustments lead to a 51% improvement in the ASVspoof2019-LA set.

3.2.5 Transformer. In the context of audio anti-spoofing, Transformer encoders are often integrated with other DL architectures, such as ResNet [295] or CNN [18, 125]. The compact convolutional Transformer (CCT) [17] is proposed by incorporating two 2D convolutional layers before Transformer encoders to enhance generalization ability. This strategy aims to extract high-level embeddings from the input spectrogram feature, rather than directly dividing the spectrogram into patches and feeding them to the Transformer. These high embeddings obtained after the convolutional layers aggregate information from all regions of the spectrogram. [117] modifies the 2D convolutional layer into 1D, accompanied by a smaller number of Transformer encoders, named OCT, to reduce overfitting. Rawformer, proposed by [139] combines SE-Res2Net with a positional aggregator before passing to Transformer encoders. This integration of Res2Net and Transformer architectures aims to effectively capture both local and global dependencies.

The conventional Transformer architecture typically focuses on the temporal domain exclusively [253]. However, recent advancements have been made to adapt the Transformer to treat the temporal and frequency dimensions equally. Zhang et al. [276] leverage both the feature matrix and its transposed version to facilitate self-attention mechanism across both temporal and frequency domains. [235] proposes a Temporal-Frequency Transformer (TFT) module, consisting of the temporal modeling branch and the frequency modeling branch in parallel. This design effectively captures long-term dependencies in both domains simultaneously.

3.2.6 Time-Delay Neural Network (TDNN). TDNN is widely recognized in tasks like speech recognition by converting acoustic signals into phonetic representations. Various efforts have been made

to utilize TDNN variants, such as ECAPA-TDNN, for spoofed audio detection [29, 34, 46, 243]. However, the performance of TDNN in antispoofing countermeasures remains suboptimal compared to their effectiveness in speaker verification.

3.2.7 Differentiable Architecture Search (DART). DART [133] introduces a dynamic detection model, enabling optimization of both architectures and parameter values based on performance on the validation set using gradient descent. The candidate operations for network building blocks include convolutional, pooling, and residual layers. To reduce computation power and memory usage, Ge et al. [70] propose a partially-connected DARTS (PC-DARTS) by adding a random mask to some partial channels during the architecture search stage. With random masking, PC-DARTS ensures complex architecture learning while reducing training time by 50% compared to standard DARTS. Building upon PC-DARTS, Raw PC-DARTS [71] further advances the methodology by leveraging a learnable SincNet filter with filter masking to handle raw input signals directly, rather than using LFCC features as the front-end. This approach leads to a 64% performance improvement in EER. In [217], Light-DARTS is introduced, incorporating the MFM module as one of the potential candidate operations within the architecture search space, where the MFM module functions as a feature selection mechanism.

3.2.8 Pooling and Attention mechanism. Pooling layers and attention mechanisms serve critical functions in back-end classifiers by highlighting discriminative information.

Statistical pooling Common pooling methods include max pooling and average pooling, which respectively select the maximum value or compute the average value along a dimension. Introduced by [198], statistics pooling, employed in x-vector architectures with TDNN, calculates and concatenates the mean and standard deviation of frame-level embeddings to generate an utterance-level representation. An enhanced version, attentive statistics pooling (ASP), proposed in [173], incorporates an attention mechanism into statistics pooling. ASP assigns channel-dependent weights to each frame, dynamically emphasizing the most informative frames during pooling. ASP has been utilized in various recent detection architectures [114, 151, 158].

Attention In recent research, attention mechanisms have been integrated into classifier architectures such as LCNN and ResNet [84, 151, 153]. Well-known attention mechanisms include convolutional block attention modules (CBAM) [236] and dual attention network (DANet)[64], which address both channel attention and spatial attention. CBAM comprises two sub-attention modules in sequence, while DANet incorporates them in parallel. Zhou et al. [297] employ topK pooling before computing attention weights across temporal and spatial dimensions, aimed at preserving a lightweight architecture. Rostami et al. [187] suggest adding a learnable attention mask to the channel dimension of the feature map.

3.2.9 Other Architectures. In addition to the commonly used architectures mentioned above, advanced techniques like quantum neural networks are also being explored to address the problem of spoofed audio. Wang et al. [222] propose employing a 4-qubit variational quantum circuit network as the back-end classifier, utilizing feature maps extracted from the pre-trained WavLM-Large model. They design a pipeline that utilizes Bi-LSTM to convert the feature maps into low-dimensional embedding vectors, which can be processed by the quantum circuit.

3.2.10 Performance discussion on classifier selection. While the existing countermeasures have shown effectiveness in detecting fully spoofed audio, each classifier type comes with its own set of limitations that require attention, as outlined in TABLE 4. Employing ensemble methods can be advantageous in addressing these limitations. For instance, CNN-based detectors are widely used for their proficiency in extracting local patterns, but they may struggle to capture long-term information. In such cases, ensembling with Transformer-based models can help in capturing global

temporal features. GATs excel in formulating inter-relationships among frame-level features, which can lead to improved performance when combined with other CNN-based detectors. Additionally, DL-based classifiers may face challenges related to overfitting, resulting in poor performance if the training dataset is small or lacks diversity.

3.3 End-to-End (E2E) Architecture

One limitation of detection models with a two-stage architecture is their high dependency on extracted features. The performance of classifiers can vary significantly depending on the chosen input features. Additionally, information lost during feature extraction is often irretrievable, such as the neglect of phase information when using power spectrum features. Therefore, E2E architectures have garnered more attention for audio anti-spoofing model development [37, 102, 267]. In E2E models, the entire process, from input to output, is encapsulated within a single network, eliminating the need for separate front-end feature processing and back-end classifier engineering stages. However, many proposed architectures claimed as E2E models actually utilize a learnable SincNet filter with fixed cutoff frequencies and Mel scales to extract a 2D spectral-temporal feature map. The use of learnable cutoff frequencies in the audio anti-spoofing task may lead to overfitting, and the pre-determined settings make them not truly representative of E2E models.

Ma et al. [155] present a genuinely E2E architecture by directly obtaining embedding from the raw waveform using a 1D convolutional layer with a kernel size of 3 followed by ResNet blocks. They also enhance the model by incorporating a 1D convolutional layer and BN layer into the skip connection of ResNet blocks, thus extending the perceptual information range. Hua et al. [87] modify the kernel size of the first 1D convolutional layer to 7 to capture longer-range dependencies.

Fang et al. [61] introduces an E2E model based on Res2Net blocks, wherein information from previous subsegments undergoes another convolutional layer with a kernel size of 1 before being added to the current subsegment. Conversely, the E2E ConvNet architecture proposed by [152] alternates between 1D convolutional layers and original Res2Net blocks, while appending a channel attention layer to the end of each Res2Net block, assigning channel-dependent weights.

3.4 Training Optimization Techniques

We evaluate a range of training optimization techniques within the domains of data augmentation, loss function, and activation function. Through a comprehensive exploration of these techniques, we aim to provide insights into their application and impact in enhancing the performance and generalization capabilities of audio anti-spoofing detection systems. TABLE 5 illustrates how integrating specific optimization techniques into detection algorithms leads to detection performance improvement.

3.4.1 Data Augmentation (DA) techniques. DA techniques serve as indispensable tools to enhance the robustness and diversity of datasets in training audio anti-spoofing models. This section assesses some commonly used DA techniques, including masking, mix-up, and codec variation, along with other DA methods.

Masking SpecAugment [176], initially introduced a DA method for speech recognition, involves randomly masking blocks of frequency bins and/or blocks of time steps on spectrograms. Over time, SpecAugment has been widely applied to improve the performance of anti-spoofing detectors [5, 33, 68, 109, 123, 257]. Additionally, [297] conducts experiments with random masking on the dimension of channels. Subsequently, SpecAverage [42] is proposed, to replace the random masking of the feature map with the average feature value rather than the value of zero.

Mix-up To prevent information loss on critical audio segments, Kim et al. [104] introduce a cut-and-mix technique, known as SpecMix. This method combines two data samples, where up

Table 5. Ablation Study of Training Optimization Techniques on the ASVspoof2019-LA, ASVspoof2021-LA, ASVspoof2021-DF Datasets

Methods	Used in literature	ASVspoof2019-LA		ASVspoof2021-LA		ASVspoof2021-DF	
		w/o	with	w/o	with	w/o	with
Data augmentation							
SpecAugment	[68]	6.51	5.139	-	-	-	-
SpecMix	[58]	-	-	4.04	3.09	-	-
Codec augmentation	[46]	-	-	19.20	9.21	-	-
	[223]	2.24	6.41	30.17	7.96	-	-
	[42]	-	-	21.41	7.22	29.31	21.60
Noise addition	[206]	-	-	9.50	5.31	-	-
	[209]	-	-	4.48	0.82	4.57	2.85
Loss function							
LMCL	[33]	4.04	3.49	-	-	-	-
OC-Softmax	[287]	4.69	2.19	-	-	-	-
SAMO	[54]	1.74	1.08	-	-	-	-
MSE	[226]	3.04	1.92	-	-	-	-
Focal loss	[46]	-	-	9.21	7.51	-	-
Center loss*	[88]	0.68	0.52	3.85	3.38	-	-

The evaluation metric is EER (%). For data augmentation, "w/o" means that no data augmentation techniques are applied, whereas "with" denotes that only the specified data augmentation method is applied. For loss function, "w/o" means that the model utilizes the CE loss with Softmax as the loss function, while "with" indicates that the loss function switches to the specified alternative. "-" indicates that the authors do not report the performance with the corresponding dataset.

* Under "with", Center loss is incorporated alongside the CE loss with Softmax.

to three frequency bins or temporal bands are masked on one sample. Then, the masked regions are replaced with features from another sample to create a new training sample. Notably, the label for the new sample is determined through a weighted average of the labels from the two original samples, with the weights assigned by the extent of the masked areas. SpecMix is effective to prevent overfitting and generalize to the unseen attack [58, 87, 213, 274]. Wang et al. [235] apply the cut-and-mix technique to training utterances under the same label. However, determining the percentage of SpecMix over the entire training set is crucial. Experimental results indicate that either entirely omitting SpecMix operations or uniformly applying SpecMix to all data can lead to performance degradation [58].

Codec augmentation This technique is utilized to replicate compression algorithms used in encoding audio signals, thereby enhancing the model's robustness to unseen coding and transmission artifacts, especially in datasets like ASVspoof2021-LA set [8, 46, 47, 49, 62, 243]. Codec augmentation can be categorized into two main types: multimedia encoding and transmission encoding. In multimedia encoding, raw audio is transformed into various codecs such as MP3, M4A, MP2, OGG, and AAC, each with different sampling rates, before being resampled back to 16kHz. Transmission encoding involves employing multiple Voice over Internet Protocol (VoIP) or telephony transformation techniques, such as G.711, G.726, Adaptive Multi-Rate Wideband (AMR-WB), and GSM, with varying bitrates. [42] simulates the random packet loss as augmentation. Meanwhile, [116, 213] apply band-pass finite impulse response (FIR) filters to mimic speech codec, which may cause information loss at specific frequency bands.

Other commonly used DA techniques include speed perturbation [223], time stretching [27], pitch shifting [27], addition of noise and room impulse responses (RIRs) [170, 206], and generation of new spoofed audio samples using various vocoders [223]. It is common practice to combine

multiple DA techniques, involving both online and offline implementations, to further enhance the model's robustness [13, 33, 115, 240]. However, experiment results indicate that the effectiveness of DA techniques may be feature-dependent or dataset-dependent. For instance, channel simulation, which includes applying codecs, adding additive noise and RIRs, does not perform well for LFCC-based anti-spoofing models [293] but is effective for models utilizing features pre-trained with wav2vec 2.0 [132].

Furthermore, DA may lead to data pollution when the augmented samples do not accurately represent the diversity of real-world conditions. To address this data pollution issue, Lin et al. [132] implement split batch normalization (SBN) within the conventional LCNN blocks, where the BN layer is split into the main branch and an auxiliary branch, allowing the main branch to process weakly augmented training data and the auxiliary branch to handle strongly augmented data concurrently. By doing so, it prevents the main branch from being affected by data pollution caused by DA during the inference stage.

3.4.2 Loss function. The selection of loss functions is also crucial to detection performance. The cross-entropy (CE) loss with Softmax is the most widely used loss function for audio anti-spoofing tasks, which consists of a fully-connected (FC) layer, the Softmax function and the cross-entropy loss. Researchers also have proposed numerous variants to address specific tasks for audio anti-spoofing. We introduce all loss functions mentioned in the literature with the formula provided.

Cross-entropy (CE) loss with Softmax Softmax is applied to generate a probability distribution over the classes. In most anti-spoofing tasks, the number of classes is typically limited to two: bona fide and spoofed. The formula of binary version of CE loss with Softmax is given as:

$$\begin{aligned}\mathcal{L}_{BCE} &= -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{\mathbf{w}_{y_i}^\top \mathbf{x}_i}}{e^{\mathbf{w}_{y_i}^\top \mathbf{x}_i} + e^{\mathbf{w}_{1-y_i}^\top \mathbf{x}_i}} \\ &= -\frac{1}{N} \sum_{i=1}^N \log(1 + e^{(\mathbf{w}_{1-y_i} - \mathbf{w}_{y_i})^\top \mathbf{x}_i}),\end{aligned}\tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \{0, 1\}$ are the feature embedding and the corresponding label, respectively. $\mathbf{w}_0, \mathbf{w}_1 \in \mathbb{R}^D$ are the weight vectors for bona fide and spoofed classes, and N is the batch size.

Additive Margin (AM)-Softmax It introduces a margin m in angular space to make both classes' embedding distributions more compact and encourage larger angular distances between feature vectors of different classes:

$$\mathcal{L}_{AM} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{\alpha(\cos(\theta_{y_i} + m))}}{e^{\alpha(\cos(\theta_{y_i} + m))} + e^{\alpha(\cos(\theta_{1-y_i}))}},\tag{2}$$

where $\cos \theta_{y_i} = \hat{\mathbf{w}}_{y_i}^\top \hat{\mathbf{x}}_i$ is the cosine distance between length normalized vector, $\hat{\mathbf{w}}, \hat{\mathbf{x}}$ are the normalized \mathbf{w} and \mathbf{x} , and α is a scale factor.

Large Margin Cosine loss (LMCL) The LMCL also aims to maximize the inter-class variance and minimize the intra-class variance. In LMCL, the margin is added in the cosine space rather than angular space, as follow:

$$\begin{aligned}\mathcal{L}_{LMCL} &= -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{\alpha(\hat{\mathbf{w}}_{y_i}^\top \hat{\mathbf{x}}_i - m)}}{e^{\alpha(\hat{\mathbf{w}}_{y_i}^\top \hat{\mathbf{x}}_i - m)} + e^{\alpha(\hat{\mathbf{w}}_{1-y_i}^\top \hat{\mathbf{x}}_i)}} \\ &= -\frac{1}{N} \sum_{i=1}^N \log(1 + e^{\alpha(m - (\hat{\mathbf{w}}_{y_i} - \hat{\mathbf{w}}_{1-y_i})^\top \hat{\mathbf{x}}_i)}).\end{aligned}\tag{3}$$

The margin term in LMCL helps make the model more robust to noisy samples.

One Class (OC)-Softmax The theoretical decision boundary of genuine speech and spoofed speech remains the same angle to the weight vectors across Softmax, AM-Softmax, and LMCL. Zhang et al. [287] point out that employing a uniform compact margin for both genuine and spoofed speech can lead to overfitting to known attacks. Therefore, OC-Softmax is proposed, which uses two different margins, m_0 and m_1 , to compact the genuine speech as well as simultaneously isolate the spoofing speech. OC-Softmax is denoted as:

$$\mathcal{L}_{OC} = -\frac{1}{N} \sum_{i=1}^N \log(1 + e^{\alpha(m_{y_i} - \hat{\mathbf{w}}_0 \hat{\mathbf{x}}_i (-1)^{y_i})}). \quad (4)$$

Several comparative studies have been conducted to demonstrate the generalizability of OC-Softmax, particularly towards unknown attacks [82, 158]. Moreover, recent efforts have concentrated on enhancing OC-Softmax even further. Ding et al. [54] propose speaker attractor multicenter one-class learning (SAMO), which aims to construct multiple clusters for bonafide utterances based on individual speakers, rather than clustering all bonafide speeches into one group. The SAMO is defined as:

$$\mathcal{L}_{SAMO} = -\frac{1}{N} \sum_{i=1}^N \log(1 + e^{\alpha(m_{y_i} - d_i)(-1)^{y_i}}), \quad (5)$$

where d_i is computed by

$$d_i = \begin{cases} \hat{\mathbf{w}}_{s_i} \hat{\mathbf{x}}_i & \text{if } y_i = 0 \\ \max_s (\hat{\mathbf{w}}_s \hat{\mathbf{x}}_i), s \in \mathcal{S}_{train} & \text{if } y_i = 1. \end{cases}$$

Here, s represents individual speaker, and $\hat{\mathbf{w}}_{s_i}$ is the normalized speaker vector. Ren et al. [185] add a dispersion loss specifically tailored for spoofed samples to the OC-Softmax. This addition aims to make the real speech space more compact than the originals, thereby encouraging known spoofing samples to cover the entire spoofing space. As a result, unknown spoofing attacks are more likely to be classified within the spoofing space.

Mean Square Error (MSE) The margin-based loss is sensitive to the hyper-parameter settings. Wang et al. [226] suggest utilizing the MSE between the model detection output and the ground truth, as a hyperparameter-free loss function:

$$\mathcal{L}_{MSE} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=0}^{C-1} (\hat{\mathbf{w}}_k^T \hat{\mathbf{x}}_i - \mathbb{1}(y_i = k))^2. \quad (6)$$

where C indicates the total number of classes.

Triplet loss Triplet loss is utilized to project audio samples into an embedding space. Samples sharing the same labels are brought closer together, determined by the Euclidean distance, while those with differing labels are distanced by a specified margin. The formula for Triplet loss is shown as Eq. 7.

$$\mathcal{L}_{Tri} = \sum_{i=1}^N \|f(\mathbf{z}_i^a) - f(\mathbf{z}_i^r)\|_2^2 - \|f(\mathbf{z}_i^a) - f(\mathbf{z}_i^s)\|_2^2 + \alpha, \quad (7)$$

where \mathbf{z}_i^a represent the anchor sample, \mathbf{z}_i^r is a positive sample of the same class as anchor sample, \mathbf{z}_i^s is a negative sample of a different class from anchor sample, $f(\cdot)$ extracts embedding representations for each sample, and α is a pre-defined margin. In the context of audio anti-spoofing, Triplet loss is typically combined with other objective functions [211, 247].

Focal loss Focal loss is another common objective function for audio anti-spoofing tasks, especially when dealing with data imbalance during the training stage. The focal loss can be

formulated by Eq. 9.

$$\mathcal{L}_{Focal} = -(1 - p_t)^\gamma \log(p_t). \quad (8)$$

It adds $(1 - p_t)^\gamma$ to the standard CE loss, where p_t is the predicted probability of the genuine class and γ controls the rate at which the loss for well-classified examples is down-weighted.

Center loss Huang et al. [88] propose to integrate Center loss alongside the CE loss to minimize intra-class variability of deep embeddings. This addition is significant because the CE loss primarily focuses on capturing inter-class differences. The definition of Center loss is indicated as follows:

$$\mathcal{L}_{Center} = \frac{1}{2} \sum_{i=1}^N \|\mathbf{w}_{y_i}^\top \mathbf{x}_i - c_{y_i}\|_2^2, \quad (9)$$

where c_{y_i} is the centroid of the class to which the i -th input data belongs.

3.4.3 Activation function. Research has explored modifying the activation functions as well. Kang et al. [98] conduct experiments with multiple E2E models and find that the learnable activation functions, such as parametric-ReLU (PReLU) [86] and Attention-ReLU (ARelu) [31], improve the performance greater than non-learnable activation functions. Additionally, ensembling both learnable and non-learnable activation functions by addition, including ReLU, ARelu, PReLU, LeakyReLU, and ELU, outperforms using any single one of the activation functions solely [97].

4 TRAINING AND ROBUSTNESS ADVANCEMENTS IN FULLY SPOOFED DETECTION

In addition to focusing on detection accuracy, numerous research efforts have been directed toward enhancing the model's robustness, efficiency, and interpretability. In the following section, we will introduce recent advances in these research directions.

4.1 Training strategies

Various training techniques have been proposed to address challenges such as model complexity, data scarcity, and computational efficiency. For instance, Xie et al. [249] leverage the Siamese network on SOTA architectures like LCNN, ResNet-18, and SE-Net, to learn a more compact and meaningful representation for audio samples without increasing the number of parameters. Notably, the Siamese network also demonstrates effectiveness in highly unbalanced datasets. Except for the Siamese network, Low-Rank Adaption and knowledge distillation are also widely adopted in audio anti-spoofing to transfer knowledge to a more efficient detection structure.

Low-Rank Adaption (LoRA) LoRA is an efficient transfer learning method that introduces two low-rank adaptive matrices specifically trained for new downstream tasks or domains [285]. Compared to traditional transfer learning methods [5, 202], LoRA provides a low-cost incremental learning process and prevents catastrophic forgetting by keeping the entire parameters of the source model unchanged. Wang et al. [219] incorporate LoRA into the multi-head attention module within the wav2vec 2.0 Transformer structure. This integration aims to produce a new set of deep embeddings better adapted to the data in various domains.

Knowledge Distillation (KD) The large-scale detection models make them difficult to deploy on edge devices with limited memory and computational power. KD is proposed to capture knowledge in a complex detection architecture into a smaller, single model that is much easier to deploy without significant loss in performance [130, 185]. Xue et al. [251] introduce a self-distillation approach, where the student model is trained to learn prediction-based knowledge through computing Kullback-Leibler (KL) divergence loss of Softmax outputs between students and teachers, as well as feature-based knowledge by computing the MSE loss between feature maps of students and teachers. Lu et al. [145] propose an offline distillation method that freezes the parameters of the teacher model to train the student model, while the student model is trained only on bonafide

data to maximize learning of the genuine speech feature space. Ren et al. [184] suggest that online distillation could enable the student model to learn extra knowledge from mutual learning with the teacher model.

4.2 Interpretability of results

Several existing works have leveraged explainable artificial intelligence (XAI) tools [9] to uncover the behaviour of deep neural network algorithms in detecting spoofed audio. Ge et al. [72] is the first effort to utilize SHapley Additive exPlanations (SHAP) [146] scores to explain detection results on the ASVspoof2019-LA dataset, by testing on speech and non-speech intervals and different subbands. The observation reveals that the classifier has learned to focus on non-speech intervals and highlighted the attention at the low-frequency sub-band at 0.5-0.6kHz. Lim et al. [131] apply both Deep Taylor [164] and layer-wise relevance propagation (LRP) [19] to learn the attribution score of audio formats in spectrograms. Furthermore, the Gradient-weighted Class Activation Mapping (Grad-CAM) [191] is used in [111, 207] to identify the significant frequency ranges in the spectrogram.

4.3 Defense to adversarial attacks

[136] has shown that the majority of anti-spoofing models are vulnerable to adversarial attacks, such as the Fast Gradient Sign Method (FGSM) [78] and Projected Gradient Descent (PGD) [156], especially when dealing with models of smaller scales. In response, various defence strategies have been proposed, primarily based on adversarial training, where anti-spoofing models are retrained using the adversarial examples generated by the PGD method [241] or the FGSM method [75, 170]. [171] claims that attackers may focus on the range beyond human perception to maximize the attack effectiveness, therefore, they suggest augmenting training data with frequency band-pass filtering and denoising to defend such attacks. Furthermore, Liao et al. [130] utilize knowledge distillation as another defence method. This is because the soft targets learned by the student models capture more nuanced information about the decision boundary, enhancing robustness against adversarial attacks.

4.4 Robustness on cross-dataset

There are two practical scenarios involving multi-dataset. One is multi-dataset co-training, where datasets from different domains are combined as training data. Co-training encourages the training models to learn information from all provided domains. The other one is cross-dataset evaluation, where models trained on a specific dataset are tested on various out-of-domain datasets. This evaluation method helps assess the robustness and generalizability of models across different data domains.

Multi-dataset co-training The experiments indicate that simply combining data from different domains does not guarantee an increase in the generalization of detection models due to domain mismatch. To address this challenge, Shim et al. [194] propose a gradient-based method that considers reducing the curvature of neighbourhoods in the loss surface while minimizing the loss function. It effectively reduces the gap between variances of multi-domain datasets. Zhang et al. [286] adopt the Regularized Adaptive Weight Modification (RAWM) to overcome the catastrophic forgetting problem caused by finetuning a trained model with an out-of-domain dataset. Wang et al. [229] use the negative energy-based certainty score [138] to evaluate the usefulness of each data in the pool consisting of datasets from various domains. They then employ the active learning technique to select the useful data to be the training data for fine-tuning.

Cross-dataset evaluation Muller et al. [165] evaluate several SOTA anti-spoofing detection models trained using the ASVspoof2019-LA dataset on the ITW dataset. The results reveal a

Table 6. The Cross-dataset Performance of Single-System State-of-the-art Models. All Models are trained or fine-tuned on ASVspoof2019-LA Training and Development Set, and evaluated on ASVspoof2019-LA Evaluation Set and In-The-Wild Dataset.

Publication	Data augmentation	Feature	Classifier	Loss function	ASVspoof 19-LA	ITW
[253]	IH&MMSec'23	Mel-Spec	Patched Transformer	CE	4.54	29.72
[218]	INTERSPEECH'23	Duration + pronunciation + wav2vec2.0-XLSR	LCNN → Bi-LSTM → MLP	CE	1.58	36.84
[248]	INTERSPEECH'23	wav2vec2.0-XLSR	LCNN → Transformer	CE, Triplet, Adversarial	0.63	24.50
[230]	ICASSP'23	wav2vec2.0-XLSR	MLP	CE	2.98	26.65
[289]	SPL'24	SpecAugment	CNN → GRU → MLP	AM-Softmax	1.79	29.66
[263]	ICASSP'24	wav2vec2.0-XLSR	ResNet-18	CE	2.07	29.19
[263]	ICASSP'24	Hubert	ResNet-18	CE	6.78	27.48
[216]	ICASSP'24	Multi-scale permutation entropy	SE-ResNet	CE	20.24	29.62
[145]*	ICASSP'24	CNN → wav2vec2.0	AASIST	CE	0.39	7.68
[231]*	ICASSP'24	Rawboost	wav2vec2.0-XLSR-Vox	MLP	0.13	12.50

The evaluation metric is EER (%). The bold values refer to the best performance on the same dataset. "+" indicates multiple techniques processed in parallel, while "→" denotes sequential order. "w/o" means that no data augmentation techniques are applied.

* [145] and [231] utilize knowledge distillation. The reported evaluation results on both datasets are produced by the student model.

significant performance decline, with some models showing random guessing behaviour. [293] suggests the performance degradation may be due to the channel effect mismatch among different datasets, as evidenced by variations in the average spectra magnitude for each dataset. To address this issue, an additional channel classifier with a Gradient Reversal Layer (GRL) is added in [293] as a discriminator, making the detecting model more robust to channel variation while preserving its discriminative power in spoofing detection. Recently, GRL has been widely used for domain adaptation in cross-dataset, and cross-language detections, encouraging models to learn domain-invariant representations [11, 299]. Besides, Salvi et al. [189] introduce a sub-network consisting of three layers of FC, dropout, BN, and LeakyReLU, operating in parallel to the detection classifier structure as a reliability estimator. This estimator evaluates each segment of the input audio, ensuring that all input segments contributing to the detection decision are both discriminative and robust enough. Segments considered insufficiently reliable are discarded from the training stage. According to Table 6, the knowledge distillation technique significantly improves the generalizability of student models in cross-dataset evaluations.

5 INTEGRATION OF ASV TO ANTI-SPOOFING COUNTERMEASURES

While existing anti-spoofing algorithms have demonstrated the effectiveness of detecting TTS or VC attacks, even under telephony or various codec conditions, the reliability of ASV systems remains vulnerable. ASV systems have been used as a biometric authentication technique, which not only requires detecting the authentication of the speech clips but also needs to verify the identity of speakers. The relationship between anti-spoofing CMs and ASV systems has been shown in Figure 1. Therefore, there is a need to integrate the ASV functionality into the anti-spoofing architectures.

The spoofing-aware speaker verification (SASV) challenge has been launched to encourage the development of single models which can detect the speech spoken by different speakers and spoofed speech [94]. Based on the challenge protocols, the VoxCeleb2 and the ASVspoof2019-LA database are used for training. The primary metric for evaluation is the SASV-EER, which treats bonafide speech from the target speaker as positive cases and all others as negative. The secondary

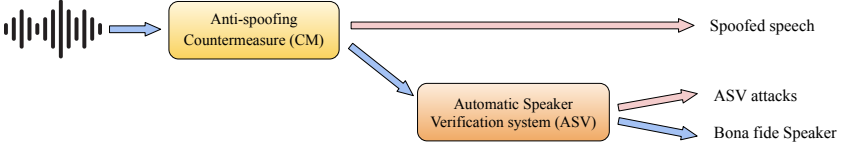


Fig. 1. Relationship of ASV systems and Anti-Spoofing CMs

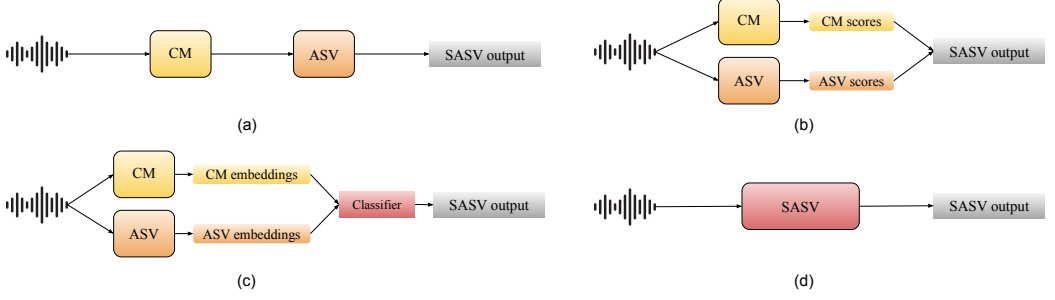


Fig. 2. Different structures of current SASV models. (a) Cascaded System (b) Score-level Fusion, (c) Embedding-level Fusion, and (d) Integrated (E2E) System.

metrics are speaker verification (SV)-EER and spoofing (SPF)-EER, which measure the capability of countermeasure and speaker verification respectively.

5.1 SOTA for SASV models

The current SASV models can be categorized into four types, as Figure 2 shows. Each category is discussed in the following.

Cascaded systems The cascaded system is the concatenation of the ASV and CM classifiers, both of which are pre-trained. Wang et al. [225] propose a cascaded system beginning with an ASV binary classifier followed by the CM detector. The ASV and CM modules are trained separately. During testing, only the test audio labelled positive by the ASV classifier will be passed to the second CM module. However, they also find that the order of the modules affects the output performance.

Score-level Fusion Score-level fusion indicates that combining the individual scores from the ASV and CM classifiers to generate an ensemble score for the SASV system. Shim et al. [195] use a score-sum fusion technique to integrate the ASV and CM sub-systems, serving as the baseline for the SASV challenge. Zhang et al. [292] introduce a probabilistic framework based on the product rule, that leverages scores from the ASV and CM subsystems, further enhancing the score-sum fusion baseline. Wu et al. [238] suggest that the baseline does not consider the disparity in scale ranges between ASV and CM scores, so they utilize a fusion technique that concatenates scores from multiple sub-CM models and sub-ASV models before passing them to the prediction layer. This approach is refined in [242], by applying the average pooling to CM embeddings to obtain a single CM score for concatenation with the rest of the sub-ASV scores. Utilizing the score-level fusion with the pooling strategy allows for flexible adjustment of the model's scale. [282] addresses the inconsistent score distribution of the CM and ASV subsystems by utilizing the L2-normalized inner product for speaker embeddings. Alenin et al. [5] use quality measurements of audio files, such as speech length and mean value of CM system scores, to penalize the ASV and CM scores. It helps to normalize target and impostor distributions.

Table 7. The Performance of State-of-the-art SASV Models on the ASVspoof2019-LA evaluation set

Publication		Category	Algorithms for ASV	Algorithms for CM	SV-EER ↓	SPF-EER ↓	SASV-EER ↓	Accessibility
[225]	INTERSPEECH'22	Cascaded	SE-ResNet-34, ECAPA-TDNN	AASIST	0.90	0.26	0.29	No
[195]	ODYSSEY'22	Score Fusion	ECAPA-TDNN	AASIST	1.66	1.76	1.71	Yes ¹
[292]	ODYSSEY'22	Score Fusion	ECAPA-TDNN	AASIST	1.94	0.80	1.53	Yes ²
[238]	ODYSSEY'22	Score Fusion	ECAPA-TDNN, ResNet-34, MFA-Comformer	AASIST, RawGAT	1.20	1.15	1.17	No
[242]	INTERSPEECH'22	Score Fusion	ECAPA-TDNN, ResNet-34, MFA-Comformer	AASIST, RawGAT	1.15	0.56	0.97	No
[282]	INTERSPEECH'22	Score Fusion	SE-Res2Net-50	AASIST	0.48	0.78	0.63	Yes ³
[5]	INTERSPEECH'22	Score Fusion	ResNet-48	ResNet-48	0.19	0.25	0.22	No
[158]	arXiv'22	Score Fusion	wav2vec 2.0, ECAPA-TDNN	AASIST	0.97	0.58	0.84	No
[195]	ODYSSEY'22	Embedding Fusion	ECAPA-TDNN	AASIST	11.48	0.78	6.37	Yes ¹
[277]	INTERSPEECH'22	Embedding Fusion	ECAPA-TDNN	AASIST	2.02	0.50	0.99	No
[290]	INTERSPEECH'22	Embedding Fusion	ECAPA-TDNN	CNN	6.24	1.73	4.78	No
[38]	INTERSPEECH'22	Embedding Fusion	Res2Net	AASIST	0.28	0.28	0.28	No
[73]	INTERSPEECH'22	Embedding Fusion	ResNet-34	AASIST	1.53	0.75	2.44	Yes ⁴
[79]	arXiv'22	Embedding Fusion	ECAPA-TDNN	AASIST, RawNet2	3.62	0.61	2.90	No
[74]	ICASSP'23	Embedding Fusion	ResNet-34	AASIST	2.34	0.80	1.49	Yes ⁴
[211]	INTERSPEECH'22	Integrated System	ECAPA-TDNN, AResNet		8.06	0.50	4.86	No
[96]	INTERSPEECH'22	Integrated System	ECAPA-TDNN		6.83	8.36	4.43	No
[167]	INTERSPEECH'23	Integrated System	MFA-Conformer		1.83	0.58	1.19	Yes ⁵

¹ https://github.com/sasv-challenge/SASVC2022_Baseline² https://github.com/zyyouzhang/SASV_PR³ https://github.com/WebPrague/SASV2022_DoubleRoc⁴ <https://github.com/eurecom-as/sasv-joint-optimisation>⁵ <https://github.com/sasv-challenge/ASVspoof5-SASVBaseline>

All models are trained using the ASVspoof2019-LA training and development set, as well as the VoxCeleb2 dataset. “↓” indicates that a lower score corresponds to better detection performance for all evaluation metrics. The bold values refer to the best performance of each subcategory

Feature Embedding-level Fusion Feature-level fusion involves either the concatenation of the ASV and CM embeddings from their extractors or the extraction of an integrated embedding to represent both ASV and CM information. Another baseline of the SASV challenge mentioned in [195] is to fuse the embeddings from the ASV and CM module together by concatenation and pass them into a simple DNN-based classifier containing three FC layers. Zhang et al. [277] apply circulant matrix transformation to ASV and CM embeddings before stacking them together. Parallel attention and SE attention are added to the back-end classifier to learn the global relationship between these two embeddings. In [290], the authors propose a “total-divide-total” structure, adding a dual-branch network to a pre-trained ASV system. After the global feature aggregation, two parallel branches extract the ASV embedding and CM embeddings separately by independent training. This model keeps the parameters of the pre-trained backbone fixed during the training

of CM branch, which reduces training time, however, the resulting performance is not quite as competitive as other SOTA. Choi et al. [38] want to reform speaker embedding to integrate the information from CM embeddings, rather than simply applying the concatenation or stacking. The FiLM [178] technique is utilized to obtain the spoofing-aware speaker embedding (SASE), conditioning on both speaker and CM embeddings through affine transformation.

In previously mentioned models, embeddings are separately optimized. However, Ge et al. [73] introduces the concept of joint optimization, by updating the ASV, CM embeddings and the back classifier simultaneously. They suggest that through joint optimization, the strength of one subsystem may compensate for the weaknesses of the other [74].

Integrated/ E2E system Mun et al. [167] propose a multi-stage training scheme on a multi-scale feature aggregation Conformer (MFA-Conformer) to obtain an SASV embedding directly rather than using separate ASV and CM models. The embedding encoder captures the mutual characteristics of ASV and CM throughout the multiple training and fine-tuning. This design not only contributes to developing a single integrated system for the SASV task but also addresses the lack of spoofed data, reducing the impact of data imbalance. However, its computational cost has yet to be discussed in the literature. In the E2E model proposed by [211], spoofed audios are labelled as TTS and VC separately, based on their generating methods, while bonafide audios receive different labels based on the speaker’s identity. The boundaries between different labels also be distinct. Kang et al. [96] modify the AM-Softmax loss by considering CM labels into account, to fine-tune the pre-trained ASV system. Liu et al. [141] experiment several domain adaptation techniques, such as Probabilistic Linear Discriminant Analysis (PLDA) [22] and CORAL [203], to transform speaker embeddings to adapt to the new domain of spoofing attacks.

5.2 Performance discussion on SASV models

Based on the same training and evaluation datasets, the performance of the SOTA methods is presented in Table 7. Currently, the SOTAs still highly rely on the capability of independent ASV and CM subsystems. Surprisingly, according to the EER metrics, simple ensemble mechanisms, such as score-fusion and cascaded systems, achieve a better performance than a single integrated SASV system. However, directly concatenating and stacking the ASV and CM embeddings reduces the ability of speaker verification, as reflected by SV-EER, even though feature-level fusion algorithms have less false alarm rate than cascaded systems. These results suggest that a single system may require a new latent space to effectively represent both speakers and spoofing information. Therefore, the future direction should focus on joint optimization and the development of an integrated SASV system to improve overall performance.

6 PARTIALLY SPOOFED DETECTION

Partially spoofed utterance can be created by inserting one or more clips of synthetic speech into the original real speech, such as changing some words within one expression. The series of the ASVspoof Challenges does not currently address this type of spoofed speech. The PF attack was initially introduced in the ADD 2022 Challenge [265], and in ADD 2023 [266], this challenge was extended to not only detect manipulated intervals but also localize the boundaries within the utterance.

Both the ADD 2022 and 2023 challenges provide testing datasets for the partially spoofed track. The primary difference in ADD2023 PF datasets, compared to ADD2022, is that ADD2023 provides the ground truth labels for both utterance level and segment level, which allows for the evaluation of accuracy in localizing fake segments within the utterance. Consequently, the primary evaluation metric for the ADD2022 PF dataset is utterance-level EER. In contrast, the key evaluation metric for the ADD2023 PF dataset is a weighted sum of utterance-level accuracy and frame-level F1 score.

Table 8. The Performance of State-of-the-art Partially Spoofed Detection Models on the evaluation set of ADD2022-PF, ADD2023-PF, and PartialSpoof datasets

Publication		Category	Feature	Classifier	PartialSpoof ↓		ADD 2022-PF↓	ADD 2023-PF↑
					Utterance- level	Segment- level		
[281]	INTERSPEECH'21	Frame-level	LFCC	LCNN-LSTM	6.19	16.21	-	-
[180]	WIFS'22	Frame-level	LFB-Spec	xResNet	10.58	-	-	-
[159]	DADA'23	Frame-level	wav2vec2.0	LSTM	-	-	-	59.62
[134]	DADA'23	Frame-level	ResNet	Bi-LSTM, CNN	-	-	-	62.49
[280]	INTERSPEECH'21	Multi-task	LFCC	SE-LCNN, LSTM	5.90	17.55	-	-
[278]	TASLP'22	Multi-task	wav2vec2.0-large	Gated-MLP	0.49	9.24	-	-
[122]	DADA'23	Multi-task	Mel-Spec	CNN, RNN	-	-	-	62.02
[118]	DADA'23	Multi-task	E2E: wav2vec2.0, AASIST		-	-	-	58.65
[239]	ICASSP'22	Boundary detection	LFCC	SE-Net, Transformer	-	-	11.1	-
[148]	ICASSP'22	Boundary detection	wav2vec2.0-XLSR	MLP, Transformer	-	-	4.80	-
[23]	ICASSP'22	Boundary detection	wav2vec2.0, ResNet	Transformer, Bi-LSTM	-	-	6.58	-
[24]	DADA'23	Boundary detection	WavLM, ResNet	Transformer, Bi-LST	-	-	-	67.13

The evaluation metric for PartialSpoof and ADD2022-PF is EER (%). The evaluation metric for ADD2023-PF dataset is a weighted sum of utterance-level accuracy and frame-level F1 score. "↓" indicates that a lower score corresponds to better detection performance, while "↑" means the opposite. "-" indicates that the authors do not report the performance with the corresponding dataset. The bold values refer to the best performance on the same dataset.

Additionally, the PartialSpoof dataset, another widely used partially spoofed dataset derived from the ASVspoof2019 set, also offers both utterance-level and segment-level ground truth and utilizes EER as the main metrics. The latest literature, evaluated using the mentioned datasets along with their corresponding metrics, is summarized and presented in TABLE 8.

6.1 SOTA models for partially spoofed detection

The SOTA for detecting partially spoofed speeches can be categorized into three main categories: frame-level detection, multi-task learning strategies, and boundary detection.

Frame-level The speech can be divided into small segments and the frame-level detection algorithms assign a genuine or spoofed label to each segment. Kumar et al. [108] utilize the GMM to calculate the log-likelihood score for each frame and assign frame-level labels. Originally designed to detect full fake audio, this system has shown potential for detecting partially spoofed speech. [180] and [281] also attempt to calculate the frame-level score by utilizing the PLDA classifier, LCNN backbone with Bi-LSTM for smoothing.

The frame-level algorithms often face the challenge of identifying small frames as spoofed within genuine segments. However, it's reasonable to expect that the manipulated part of the attacked speech should be longer than just a few frames, especially when using a small frame size. For instance, the spoofed segment should ideally be no shorter than the duration of a phoneme. Zhang et al. [275] propose a swap algorithm to switch labels when only one fake frame occurs between real labels in the sequence, and vice versa. This approach obtains longer spoofed segments. [159] suggests that the wav2vec 2.0 SSL-based feature tends to detect artifacts on the boundaries in the inserted clips as spoofed frames. Therefore, in their post-processing strategies, spoofed frames work as boundary indicators. For example, if two non-consecutive short segments are identified as spoofed, the intermediate segments are also labelled as spoofed. If only one short segment is

detected as fake, acting as a boundary between two segments, the longer segment is labelled as genuine and the shorter one is labelled as spoofed. Liu et al. [134] introduce an isolated-frame penalty term in the loss function to deal with outliers. Differences between each frame and its surrounding frames are calculated, and these differences are then summed up to form a regularity constraint in the loss function.

Multi-task In addition to the frame-level labels, utterance-level labels are also considered during the training process, as multi-task learning. Zhang et al. [280] first implement multi-task learning frameworks to construct a binary-branch structure, where frame-level and utterance-level tasks have individual classifiers and loss functions after the jointly embedding encoder. Li et al. [122] also utilize a fused loss function with a sum of frame-level and clip-level CE loss, where the utterance-level label is calculated by the weighted average on the frame-level labels. Instead of employing a fused loss function for both frame-level and utterance-level detection, [118] integrates two back-end models, ASSIST and wav2vec 2.0. This decision is based on the authors' observation that ASSIST tends to misidentify spoofed segments, while wav2vec 2.0 exhibits a bias towards the real class. Therefore, the proposed model combines the detection results from ASSIST at the utterance level and wav2vec 2.0 at the frame level.

Boundary detection The method of boundary detection aims to identify the transition boundaries between genuine and spoofed segments. Wu et al. [239] add one FC layer as the question-answer (QA) layer to assess each frame's potential as the start or end position of the fake clips. Softmax is applied as a QA loss function, and the QA loss and CE loss are summed up for training. The model proposed by [239] achieved second place in the PF track of the ADD 2022 Challenge.

Utilizing boundary detection can address post-processing needs in frame-level score-based algorithms. Cai et al. [24] train two systems with identical architecture for the task of boundary detection and frame-level detection respectively. The scores from the boundary detection system are utilized as a reference to determine the outline frame labels. The proposed model achieves the first rank in the PF track of ADD 2023.

6.2 Performance discussion on partially spoofed detection

Recent research has reached a consensus that the manipulations of PF speech mostly occur in the time domain, resulting in the presence of more artifacts. As a result, incorporating a RNN architecture [122, 159, 281] into the detector structure is preferred to leverage temporal information. RNNs can also provide global information to CNNs, which are limited by their fixed receptive region. Additionally, SSL features with fine-tuning [140, 149, 237, 278] are widely utilized in PF detection due to their effectiveness in capturing temporal domain characteristics.

Manipulated segments in PF utterances can be either spoofed segments or some arbitrary real segments from other speakers. Some work has been done to address the problem of speaker verification in the PF case. For example, the cluster module in [159] and the Variational Autoencoder module [24] work as a supplementary speaker verification system to detect outliers. However, the current testing datasets have not included scenarios involving real segments from other speakers, making it challenging to verify the effectiveness of these implementations at this stage.

7 CURRENT CHALLENGE AND FUTURE WORK

Aside from the efforts of all presented works in improving detection performance, generalization, interpretability, and robustness, there are certain limitations in the current work that need to be addressed. This section summarizes the challenges observed in the review and provides future directions for audio anti-spoofing detection development.

Reproducibility of detection models Only around 10% of journal or conference papers in the field of audio anti-spoofing offer source code for replication. Particularly concerning is the absence

of vital information related to model training, such as loss functions and hyperparameter configurations, in some publications. Enhancing reproducibility is essential for advancing research in the future and ensuring the reliability of findings. The transparency of detection models encourages researchers to build upon previous work and to improve the quality of models by identifying and eliminating errors, inconsistencies, or biases in prior research.

Diversity of available datasets The predominant research lies on datasets in the ASVspoof series. Even though more new datasets like ITW, and WaveFake have been developing, there remains a substantial gap between these experimental datasets and the realistic conditions encountered in daily life. Therefore, there is an emerging need for developing real-world datasets that encompass a wide range of conditions, including speaker diversity, spoofing generation techniques, transmission channels, environmental factors, sound effects, noise distortion, and language variations, for both fully and partially spoofed scenarios.

Cross-dataset generalization Along with the development of dataset diversity, the anti-spoofing models should gain the capability to generalize effectively across multiple datasets and domains. Leveraging transfer learning and domain adaptation techniques represents a promising direction as future work for improving the generalization on unseen spoofing attacks, audio conditions and diverse languages while preserving the discriminative power on the known conditions.

Interpretability of detection results The current stage of exploring the interpretability of detection results is typically achieved by applying various XAI tools to the trained models. However, it is also nontrivial to consider both detection performance and outcome concurrently in the design of detection architectures. Employing techniques like attention mechanisms, and feature visualization provides insights into the underlying decision-making process of spoofing detections. Improving the interpretability of detection results ultimately enhances trust in audio anti-spoofing systems.

Robustness to adversarial attacks Most of the existing approaches to defending adversarial attacks depend on adversarial training, which generates adversarial samples on known attacks to retrain the model, requiring expensive computational costs. Future directions can be deploying generator-and-discriminator mechanisms to learn from domain-invariant attacks. Furthermore, investigating adversarial attacks across different modalities may uncover potential vulnerabilities and enhance the resilience of current models.

Streaming / Real-time detection Not much research has worked on real-time detection, including fake phone calls, IoT edge devices or other low-latency conditions. It requires detection models to be computationally efficient. Techniques like model pruning, distributed computing, and real-time incremental learning can be integrated into audio anti-spoofing systems.

Privacy preservation The development and deployment of real-time anti-spoofing detection systems on smart devices have the potential to raise privacy concerns regarding access to users' biometric data and model parameters. To address these concerns, privacy-preserving techniques like secure multiparty computation can be incorporated to limit access from both the client and server sides.

Detection source tracing It is nontrivial to trace and identify the spoofing tools or audio fabricating algorithms while developing anti-spoofing technologies. By categorizing attributes associated with spoofing attacks, we not only enhance our understanding of spoofing system architecture but also generate potential forensic evidence to preserve the trustworthiness of anti-spoofing decisions.

8 CONCLUSION

In conclusion, this survey reviews the advanced audio anti-spoofing algorithms, including the key aspects of model architecture, training techniques, application scenarios, and available datasets.

Although there are many surveys about audio anti-spoofing, they mostly focus on listing out all proposed model architectures without considering the intrinsic connection from other elements in the detection pipeline. This survey is the first one, which provides a comprehensive review of all stages of audio anti-spoofing. We present a detailed comparison and discussion of current advances in feature engineering and classifier design across diverse detection applications. Additionally, we evaluate the effectiveness of optimization techniques applied in the model training process, including data augmentation techniques, activation functions and loss functions. We provide the performance evaluation and open-source information of state-of-the-art, which fosters strong baseline selections in future experiments. Lastly, we analyze the current challenges and summarize some promising research directions for future work. We hope this survey serves as a guide and gives insight into future development for preventing malicious spoofed audio in depth.

REFERENCES

- [1] P Abdzadeh and Hadi Veisi. 2023. A Comparison of CQT Spectrogram with STFT-based Acoustic Features in Deep Learning-based Synthetic Speech Detection. *Journal of AI and Data Mining* 11, 1 (2023), 119–129.
- [2] Jahangir Alam, Abderrahim Fathan, and Woo Hyun Kang. 2021. End-to-end voice spoofing detection employing time delay neural networks and higher order statistics. In *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings* 23. Springer, 14–25.
- [3] Ehab A AlBadawy, Siwei Lyu, and Hany Farid. 2019. Detecting AI-Synthesized Speech Using Bispectral Analysis.. In *CVPR workshops*. 104–109.
- [4] Federico Alegre, Asmaa Amehraye, and Nicholas Evans. 2013. A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 1–8.
- [5] Alexander Alenin, Nikita Torgashov, Anton Okhotnikov, Rostislav Makarov, and Ivan Yakovlev. 2022. A Subnetwork Approach for Spoofing Aware Speaker Verification.. In *INTERSPEECH*. 2888–2892.
- [6] Zaynab Almutairi and Hebah Elgibreen. 2022. A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms* 15, 5 (2022), 155.
- [7] Moustafa Alzantot, Ziqi Wang, and Mani B. Srivastava. 2019. Deep Residual Neural Networks for Audio Spoofing Detection. In *Proc. Interspeech 2019*. 1078–1082. <https://doi.org/10.21437/Interspeech.2019-3174>
- [8] Rohit Arora, Anmol Arora, and Rohit Singh Rathore. 2021. Impact of Channel Variation on One-Class Learning for Spoof Detection. *arXiv preprint arXiv:2109.14900* (2021).
- [9] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bannetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.
- [10] Luigi Attorresi, Davide Salvi, Clara Borrelli, Paolo Bestagini, and Stefano Tubaro. 2022. Combining automatic speaker verification and prosody analysis for synthetic speech detection. In *International Conference on Pattern Recognition*. Springer, 247–263.
- [11] Zhongjie Ba, Qing Wen, Peng Cheng, Yuwei Wang, Feng Lin, Li Lu, and Zhenguang Liu. 2023. Transferring audio deepfake detection capability across languages. In *Proceedings of the ACM Web Conference 2023*. 2033–2044.
- [12] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296* (2021).
- [13] Naseem Babu, Prattipati Kumar, Jimson Mathew, and Udit Satija. 2022. Exploration of Bonafide and Spoofed Audio Classification Using Machine Learning Models. In *2022 IEEE 19th India Council International Conference (INDICON)*. IEEE, 1–6.
- [14] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.
- [15] BT Balamurali, Kinwah Edward Lin, Simon Lui, Jer-Ming Chen, and Dorien Herremans. 2019. Toward robust audio spoofing detection: A detailed comparison of traditional and learned features. *IEEE Access* 7 (2019), 84229–84241.
- [16] Emily R. Bartusiak and Edward J. Delp. 2021. Frequency domain-based detection of generated audio. *Electronic Imaging* 33, 4 (Jan 2021). <https://doi.org/10.2352/issn.2470-1173.2021.4.mwsf-273>
- [17] Emily R Bartusiak and Edward J Delp. 2021. Synthesized speech detection using convolutional transformer-based spectrogram analysis. In *2021 55th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 1426–1430.

- [18] Emily R Bartusiak and Edward J Delp. 2022. Transformer-based speech synthesizer attribution in an open set scenario. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 329–336.
- [19] Homanga Bharadhwaj. 2018. Layer-wise relevance propagation for explainable deep learning based speech recognition. In *2018 IEEE International symposium on signal processing and information technology (ISSPIT)*. IEEE, 168–174.
- [20] Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O'Dell, Kevin Butler, and Patrick Traynor. 2022. Who are you (i really wanna know)? detecting audio {DeepFakes} through vocal tract reconstruction. In *31st USENIX Security Symposium (USENIX Security 22)*. 2691–2708.
- [21] 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 (2021), 1–14.
- [22] Pierre-Michel Bousquet and Mickael Rouvier. 2019. On robustness of unsupervised domain adaptation for speaker recognition. In *Interspeech*.
- [23] Zexin Cai, Weiqing Wang, and Ming Li. 2023. Waveform boundary detection for partially spoofed audio. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Jun 2023). <https://doi.org/10.1109/icassp49357.2023.10094774>
- [24] Zexin Cai, Weiqing Wang, Yikang Wang, and Ming Li. 2023. The DKU-DUKEECE System for the Manipulation Region Location Task of ADD 2023. *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023)* (2023).
- [25] Diego Castan, Md Hafizur Rahman, Sarah Bakst, Chris Cobo-Kroenke, Mitchell McLaren, Martin Graciarena, and Aaron Lawson. 2022. Speaker-targeted synthetic speech detection. *The Speaker and Language Recognition Workshop (Odyssey 2022)* (Jun 2022). <https://doi.org/10.21437/odyssey.2022-9>
- [26] Nidhi Chakravarty and Mohit Dua. 2022. Noise robust ASV spoof detection using integrated features and time delay neural network. *SN Computer Science* 4, 2 (2022), 127.
- [27] Nidhi Chakravarty and Mohit Dua. 2023. Data augmentation and hybrid feature amalgamation to detect audio deep fake attacks. *Physica Scripta* 98, 9 (2023), 096001.
- [28] Sandipan Chakroborty, Anindya Roy, and Goutam Saha. 2008. Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks. *International Journal of Electronics and Communication Engineering* 2, 11 (2008), 2554–2561.
- [29] Su-Yu Chang, Kai-Cheng Wu, and Chia-Ping Chen. 2019. Transfer-representation learning for detecting spoofing attacks with converted and synthesized speech in automatic speaker verification system. *Interspeech 2019* (Sep 2019). <https://doi.org/10.21437/interspeech.2019-2014>
- [30] Chen Chen, Yaozu Song, Bohan Dai, and Deyun Chen. 2023. Twice attention networks for synthetic speech detection. *Neurocomputing* 559 (2023), 126799.
- [31] Dengsheng Chen, Jun Li, and Kai Xu. 2020. Arelu: Attention-based rectified linear unit. *arXiv preprint arXiv:2006.13858* (2020).
- [32] Feng Chen, Shiwen Deng, Tieran Zheng, Yongjun He, and Jiqing Han. 2023. Graph-based spectro-temporal dependency modeling for anti-spoofing. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [33] Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, and Elie Khoury. 2020. Generalization of Audio Deepfake Detection.. In *Odyssey*. 132–137.
- [34] Xinhui Chen, You Zhang, Ge Zhu, and Zhiyao Duan. 2021. Ur channel-robust synthetic speech detection system for ASVSPOOF 2021. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge* (Sep 2021). <https://doi.org/10.21437/asvspoof.2021-12>
- [35] Bhusan Chettri, Emmanouil Benetos, and Bob LT Sturm. 2020. Dataset artefacts in anti-spoofing systems: a case study on the ASVspoof 2017 benchmark. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 3018–3028.
- [36] K. W. Cheuk, H. Anderson, K. Agres, and D. Herremans. 2020. nnAudio: An on-the-Fly GPU Audio to Spectrogram Conversion Toolbox Using 1D Convolutional Neural Networks. *IEEE Access* 8 (2020), 161981–162003. <https://doi.org/10.1109/ACCESS.2020.3019084>
- [37] Akash Chinttha, Bao Thai, Saniat Javid Sohrawardi, Kartavya Bhatt, Andrea Hickerson, Matthew Wright, and Raymond Ptucha. 2020. Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE Journal of Selected Topics in Signal Processing* 14, 5 (2020), 1024–1037.
- [38] Jeong-Hwan Choi, Joon-Young Yang, Ye-Rin Jeoung, and Joon-Hyuk Chang. 2022. Hyu submission for the SASV challenge 2022: Reforming speaker embeddings with spoofing-aware conditioning. *Interspeech 2022* (Sep 2022). <https://doi.org/10.21437/interspeech.2022-210>
- [39] Sunmook Choi, Il-Youp Kwak, and Seungsang Oh. 2022. Overlapped Frequency-distributed network: Frequency-Aware Voice spoofing countermeasure. *Interspeech 2022* (Sep 2022). <https://doi.org/10.21437/interspeech.2022-657>

- [40] Sunmook Choi, Seungsang Oh, Jonghoon Yang, Yerin Lee, and Il-Youp Kwak. 2022. Light-weight Frequency Information Aware Neural Network Architecture for Voice Spoofing Detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 477–483.
- [41] Joon Son Chung, Arsha Nagrani, and Andrew Senior. 2018. VoxCeleb2: Deep Speaker Recognition. In *Proc. Interspeech 2018*. 1086–1090. <https://doi.org/10.21437/Interspeech.2018-1929>
- [42] Ariel Cohen, Inbal Rimon, Eran Aflalo, and Haim H Permuter. 2022. A study on data augmentation in voice anti-spoofing. *Speech Communication* 141 (2022), 56–67.
- [43] Emanuele Conti, Davide Salvi, Clara Borrelli, Brian Hosler, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Matthew C Stamm, and Stefano Tubaro. 2022. Deepfake speech detection through emotion recognition: a semantic approach. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8962–8966.
- [44] Luca Cuccovillo, Christoforos Papastergiopoulos, Anastasios Vafeiadis, Artem Yaroshchuk, Patrick Aichroth, Konstantinos Votis, and Dimitrios Tzovaras. 2022. Open challenges in synthetic speech detection. In *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 1–6.
- [45] 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.
- [46] Joaquín Cáceres, Roberto Font, Teresa Grau, and Javier Molina. 2021. The biometric Vox system for the ASVSPOOF 2021 challenge. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge* (Sep 2021). <https://doi.org/10.21437/asvspoof.2021-11>
- [47] Rohan Kumar Das. 2021. Known-unknown data augmentation strategies for detection of logical access, physical access and speech deepfake attacks: ASVspoof 2021. *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge* (2021), 29–36.
- [48] Rohan Kumar Das, Jichen Yang, and Haizhou Li. 2019. Long range acoustic and deep features perspective on ASVspoof 2019. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 1018–1025.
- [49] Rohan Kumar Das, Jichen Yang, and Haizhou Li. 2021. Data augmentation with signal companding for detection of logical access attacks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6349–6353.
- [50] Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing* 28, 4 (1980), 357–366.
- [51] Hussain Dawood, Sajid Saleem, Farman Hassan, and Ali Javed. 2022. A robust voice spoofing detection system using novel CLS-LBP features and LSTM. *Journal of King Saud University-Computer and Information Sciences* 34, 9 (2022), 7300–7312.
- [52] Jiacheng Deng, Terui Mao, Diqun Yan, Li Dong, and Mingyu Dong. 2022. Detection of synthetic speech based on spectrum defects. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. 3–8.
- [53] Hira Dhamyal, Ayesha Ali, Ihsan Ayyub Qazi, and Agha Ali Raza. 2021. Fake Audio Detection in Resource-Constrained Settings Using Microfeatures. In *Interspeech*. 4149–4153.
- [54] Siwen Ding, You Zhang, and Zhiyao Duan. 2023. Samo: Speaker attractor multi-center one-class learning for voice anti-spoofing. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [55] Abhishek Dixit, Nirmal Kaur, and Staffy Kingra. 2023. Review of audio deepfake detection techniques: Issues and prospects. *Expert Systems* 40, 8 (2023), e13322.
- [56] Thien Phuc Doan, Kihun Hong, and Souhwan Jung. 2023. GAN Discriminator based Audio Deepfake Detection. In *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes*. 29–32.
- [57] Thien-Phuc Doan, Long Nguyen-Vu, Souhwan Jung, and Kihun Hong. 2023. Bts-e: Audio deepfake detection using breathing-talking-silence encoder. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [58] Shunbo Dong, Jun Xue, Cunhang Fan, Kang Zhu, Yujie Chen, and Zhao Lv. 2023. Multi-perspective Information Fusion Res2Net with RandomSpecmix for Fake Speech Detection. *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023)* (2023).
- [59] Youngsik Eom, Yeonghyeon Lee, Ji Sub Um, and Hoi Rin Kim. 2022. Anti-spoofing using transfer learning with variational information bottleneck. *Interspeech 2022* (Sep 2022). <https://doi.org/10.21437/interspeech.2022-10200>
- [60] Cunhang Fan, Jun Xue, Shunbo Dong, Mingming Ding, Jiangyan Yi, Jinpeng Li, and Zhao Lv. 2023. Subband fusion of complex spectrogram for fake speech detection. *Speech Communication* 155 (Nov 2023), 102988. <https://doi.org/10.1016/j.specom.2023.102988>
- [61] Xin Fang, Haijia Du, Tian Gao, Liang Zou, and Zhenhua Ling. 2021. Voice spoofing detection with raw waveform based on Dual Path Res2net. In *5th International Conference on Crowd Science and Engineering*. 160–165.

- [62] Abderrahim Fathan, Jahangir Alam, and Woo Hyun Kang. 2022. Mel-spectrogram image-based end-to-end audio deepfake detection under channel-mismatched conditions. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.
- [63] Joel Frank and Lea Schönherr. 2021. Wavefake: A data set to facilitate audio deepfake detection. *arXiv preprint arXiv:2111.02813* (2021).
- [64] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3146–3154.
- [65] Quchen Fu, Zhongwei Teng, Jules White, Maria E Powell, and Douglas C Schmidt. 2022. Fastaudio: A learnable audio front-end for spoof speech detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3693–3697.
- [66] Sadaaki Furui. 1981. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29, 2 (1981), 254–272.
- [67] Yang Gao, Tyler Vuong, Mahsa Elyasi, Gaurav Bharaj, and Rita Singh. 2021. Generalized Spoofing Detection Inspired from Audio Generation Artifacts. In *Proc. Interspeech 2021*. 4184–4188. <https://doi.org/10.21437/Interspeech.2021-1705>
- [68] Yang Gao, Tyler Vuong, Mahsa Elyasi, Gaurav Bharaj, and Rita Singh. 2021. Generalized spoofing detection inspired from audio generation artifacts. *Interspeech 2021* (Aug 2021). <https://doi.org/10.21437/interspeech.2021-1705>
- [69] Konstantin Gasenzer and Moritz Wolter. 2023. Towards generalizing deep-audio fake detection networks. *arXiv preprint arXiv:2305.13033* (2023).
- [70] Wanying Ge, Michele Panariello, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2021. Partially-connected differentiable architecture search for Deepfake and spoofing detection. *Interspeech 2021* (Aug 2021). <https://doi.org/10.21437/interspeech.2021-1187>
- [71] Wanying Ge, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2021. Raw differentiable architecture search for speech deepfake and spoofing detection. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge* (Sep 2021). <https://doi.org/10.21437/asvspoof.2021-4>
- [72] Wanying Ge, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2022. Explaining deep learning models for spoofing and deepfake detection with SHapley Additive exPlanations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6387–6391.
- [73] Wanying Ge, Hemlata Tak, Massimiliano Todisco, and Nicholas Evans. 2022. On the potential of jointly-optimised solutions to spoofing attack detection and automatic speaker verification. *IberSPEECH 2022* (Nov 2022). <https://doi.org/10.21437/iberspeech.2022-11>
- [74] Wanying Ge, Hemlata Tak, Massimiliano Todisco, and Nicholas Evans. 2023. Can spoofing countermeasure and speaker verification systems be jointly optimised? *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Jun 2023). <https://doi.org/10.1109/icassp49357.2023.10095068>
- [75] Wanying Ge, Xin Wang, Junichi Yamagishi, Massimiliano Todisco, and Nicholas Evans. 2024. Spoofing attack augmentation: can differently-trained attack models improve generalisation?. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 12531–12535.
- [76] Lazaro J Gonzalez-Soler, Marta Gomez-Barrero, Madhu Kamble, Massimiliano Todisco, and Christoph Busch. 2022. Dual-stream temporal convolutional neural network for voice presentation attack detection. In *2022 International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 1–6.
- [77] Lazaro J Gonzalez-Soler, Jose Patino, Marta Gomez-Barrero, Massimiliano Todisco, Christoph Busch, and Nicholas Evans. 2020. Texture-based presentation attack detection for automatic speaker verification. In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 1–6.
- [78] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).
- [79] Petr Grinberg and Vladislav Shikhov. 2022. A Comparative Study of Fusion Methods for SASV Challenge 2022. *arXiv preprint arXiv:2203.16970* (2022).
- [80] Aswin Sankesh GS, V Ganeshkumar, Vetrivel Chelian Thirumavalavan, Velmurugan PG Sivabalan, Madhan Nanchan Suresh, and Thiruvengadam S Jayaraman. 2022. Synthetic speech classification using bidirectional LSTM Networks. In *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)*. IEEE, 1–6.
- [81] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100* (2020).
- [82] Jinlin Guo, Yancheng Zhao, and Haoran Wang. 2023. Generalized Spoof Detection and Incremental Algorithm Recognition for Voice Spoofing. *Applied Sciences* 13, 13 (2023), 7773.
- [83] Priyanka Gupta, Piyushkumar K Chodingala, and Hemant A Patil. 2022. Significance of Quadrature and In-Phase Components for Synthetic Spoofed Speech Detection. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 1252–1258.

- [84] John H.L. Hansen and ZHENYU WANG. 2022. Audio anti-spoofing using simple attention module and joint optimization based on additive angular margin loss and meta-learning. *Interspeech 2022* (Sep 2022). <https://doi.org/10.21437/interspeech.2022-904>
- [85] Taufiq Hasan, Seyed Omid Sadjadi, Gang Liu, Navid Shokouhi, Hynek Bořil, and John HL Hansen. 2013. CRSS systems for 2012 NIST speaker recognition evaluation. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 6783–6787.
- [86] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.
- [87] Guang Hua, Andrew Beng Jin Teoh, and Haijian Zhang. 2021. Towards end-to-end synthetic speech detection. *IEEE Signal Processing Letters* 28 (2021), 1265–1269.
- [88] 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.
- [89] Sundas Ibrar, Ali Javed, and Hafsa Ilyas. 2023. Voice presentation attacks detection using acoustic MLTP Features and BiLSTM. In *2023 International Conference on Communication, Computing and Digital Systems (C-CODE)*. IEEE, 1–5.
- [90] Keith Ito and Linda Johnson. 2017. The Ij speech dataset. (2017).
- [91] Zhe Ji, Zhi-Yi Li, Peng Li, Maobo An, Shengxiang Gao, Dan Wu, and Faru Zhao. 2017. Ensemble Learning for Countermeasure of Audio Replay Spoofing Attack in ASVspoof2017. In *Proc. Interspeech 2017*. 87–91. <https://doi.org/10.21437/Interspeech.2017-1246>
- [92] 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.
- [93] Jee-weon Jung, Hye-jin Shim, Hee-Soo Heo, and Ha-Jin Yu. 2019. Replay attack detection with complementary high-resolution information using end-to-end DNN for the ASVSPOOF 2019 challenge. *Interspeech 2019* (Sep 2019). <https://doi.org/10.21437/interspeech.2019-1991>
- [94] Jee-weon Jung, Hemlata Tak, Hye-jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen. 2022. SASV 2022: The first spoofing-aware speaker verification challenge. *arXiv preprint arXiv:2203.14732* (2022).
- [95] Megha Kandari, Vikas Tripathi, and Bhaskar Pant. 2023. A Comprehensive Review of Media Forensics and Deepfake Detection Technique. In *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 392–395.
- [96] Woohyun Kang, Md Jahangir Alam, and Abderrahim Fathan. 2022. End-to-end framework for spoof-aware speaker verification. *Interspeech 2022* (Sep 2022). <https://doi.org/10.21437/interspeech.2022-139>
- [97] Woo Hyun Kang, Jahangir Alam, and Abderrahim Fathan. 2021. Investigation on activation functions for robust end-to-end spoofing attack detection system. *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge* (2021), 83–88.
- [98] Woo Hyun Kang, Jahangir Alam, and Abderrahim Fathan. 2022. Attentive activation function for improving end-to-end spoofing countermeasure systems. *arXiv preprint arXiv:2205.01528* (2022).
- [99] Piotr Kawa, Marcin Plata, Michał Czuba, Piotr Szymański, and Piotr Syga. 2023. Improved deepfake detection using whisper features. *INTERSPEECH 2023* (Aug 2023). <https://doi.org/10.21437/interspeech.2023-1537>
- [100] Awais Khan and Khalid Mahmood Malik. 2023. SpotNet: A spoofing-aware Transformer Network for Effective Synthetic Speech Detection. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*. 10–18.
- [101] Awais Khan, Khalid Mahmood Malik, and Shah Nawaz. 2024. Frame-to-Utterance Convergence: A Spectra-Temporal Approach for Unified Spoofing Detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 10761–10765.
- [102] Awais Khan, Khalid Mahmood Malik, James Ryan, and Mikul Saravanan. 2023. Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures. *Artificial Intelligence Review* 56, Suppl 1 (2023), 513–566.
- [103] Zahra Khanjani, Gabrielle Watson, and Vandana P Janeja. 2021. How deep are the fakes? focusing on audio deepfake: A survey. *arXiv preprint arXiv:2111.14203* (2021).
- [104] Gwantae Kim, David K. Han, and Hanseok Ko. 2021. SpecMix : A mixed sample data augmentation method for training with time-frequency domain features. *Interspeech 2021* (Aug 2021). <https://doi.org/10.21437/interspeech.2021-103>
- [105] Juntae Kim and Sung Min Ban. 2023. Phase-aware spoof speech detection based on Res2Net with phase network. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [106] Tomi Kinnunen, Héctor Delgado, Nicholas Evans, Kong Aik Lee, Ville Vestman, Andreas Nautsch, Massimiliano Todisco, Xin Wang, Md Sahidullah, Junichi Yamagishi, et al. 2020. Tandem assessment of spoofing countermeasures

- and automatic speaker verification: Fundamentals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2195–2210.
- [107] Zvi Kons, Slava Shechtman, Alex Sorin, Carmel Rabinovitz, and Ron Hoory. 2019. High Quality, Lightweight and Adaptable TTS Using LPCNet. In *Proc. Interspeech 2019*. 176–180. <https://doi.org/10.21437/Interspeech.2019-1705>
 - [108] A Kishore Kumar, Dipjyoti Paul, Monisankha Pal, Md Sahidullah, and Goutam Saha. 2021. Speech frame selection for spoofing detection with an application to partially spoofed audio-data. *International Journal of Speech Technology* 24 (2021), 193–203.
 - [109] Il-Youp Kwak, Sunmook Choi, Jonghoon Yang, Yerin Lee, Soyul Han, and Seungsang Oh. 2022. Low-quality fake audio detection through frequency feature masking. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. 9–17.
 - [110] Il-Youp Kwak, Sungsu Kwag, Junhee Lee, Jun Ho Huh, Choong-Hoon Lee, Youngbae Jeon, Jeonghwan Hwang, and Ji Won Yoon. 2021. ResMax: Detecting voice spoofing attacks with residual network and max feature map. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 4837–4844.
 - [111] Il-Youp Kwak, Sungsu Kwag, Junhee Lee, Youngbae Jeon, Jeonghwan Hwang, Hyo-Jung Choi, Jong-Hoon Yang, So-Yul Han, Jun Ho Huh, Choong-Hoon Lee, et al. 2023. Voice spoofing detection through residual network, max feature map, and depthwise separable convolution. *IEEE Access* (2023).
 - [112] Cheng-I Lai, Nanxin Chen, Jesús Villalba, and Najim Dehak. 2019. Assert: Anti-spoofing with squeeze-excitation and residual networks. *Interspeech 2019* (Sep 2019). <https://doi.org/10.21437/interspeech.2019-1794>
 - [113] Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov. 2019. STC antispoofing systems for the ASVSPOOF2019 challenge. *Interspeech 2019* (Sep 2019). <https://doi.org/10.21437/interspeech.2019-1768>
 - [114] Jin Woo Lee, Eungbeom Kim, Junghyun Koo, and Kyogu Lee. 2022. Representation selective self-distillation and WAV2VEC 2.0 feature exploration for spoof-aware speaker verification. *Interspeech 2022* (Sep 2022). <https://doi.org/10.21437/interspeech.2022-11460>
 - [115] Yerin Lee, Narin Kim, Jaehong Jeong, and Il-Youp Kwak. 2023. Experimental Case Study of Self-Supervised Learning for Voice Spoofing Detection. *IEEE Access* 11 (2023), 24216–24226.
 - [116] Changtao Li, Feiran Yang, and Jun Yang. [n. d.]. Multi-Scale Information Aggregation for Spoofing Detection. *Available at SSRN 4251042* ([n. d.]).
 - [117] Changtao Li, Feiran Yang, and Jun Yang. 2022. The role of long-term dependency in synthetic speech detection. *IEEE Signal Processing Letters* 29 (2022), 1142–1146.
 - [118] Jun Li, Lin Li, Mengjie Luo, Xiaoqin Wang, Shushan Qiao, and Yumei Zhou. 2023. Multi-grained Backend Fusion for Manipulation Region Location of Partially Fake Audio. In *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*, Vol. 755.
 - [119] Jialong Li, Hongxia Wang, Peisong He, Sani M Abdullahi, and Bin Li. 2022. Long-term variable Q transform: A novel time-frequency transform algorithm for synthetic speech detection. *Digital Signal Processing* 120 (2022), 103256.
 - [120] Kai Li, Xugang Lu, Masato Akagi, and Masashi Unoki. 2023. Contributions of Jitter and Shimmer in the Voice for Fake Audio Detection. *IEEE Access* (2023).
 - [121] Kai Li, Yao Wang, Minh Le Nguyen, Masato Akagi, and Masashi Unoki. 2022. Analysis of amplitude and frequency perturbation in the voice for fake audio detection. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 929–936.
 - [122] Kang Li, Xiao-Min Zeng, Jian-Tao Zhang, and Yan Song. 2023. Convolutional Recurrent Neural Network and Multitask Learning for Manipulation Region Location. In *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*, Vol. 750.
 - [123] Lanting Li, Tianliang Lu, Xingbang Ma, Mengjiao Yuan, and Da Wan. 2023. Voice Deepfake Detection Using the Self-Supervised Pre-Training Model HuBERT. *Applied Sciences* 13, 14 (2023), 8488.
 - [124] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. 2022. A comparative study on physical and perceptual features for deepfake audio detection. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. 35–41.
 - [125] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. 2023. Robust Deepfake Audio Detection via Bi-Level Optimization. In *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 1–6.
 - [126] Menglu Li and Xiao-Ping Zhang. 2023. Robust Audio Anti-Spoofing System Based on Low-Frequency Sub-Band Information. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 1–5.
 - [127] Wei Li, Jichen Yang, and Pei Lin. 2023. Investigation of the influence of blocks on the linear spectrum for synthetic speech detection. *Electronics Letters* 59, 9 (2023), e12797.
 - [128] 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.

- [129] Xu Li, Xixin Wu, Hui Lu, Xunying Liu, and Helen Meng. 2021. Channel-wise gated res2net: Towards robust detection of synthetic speech attacks. *arXiv preprint arXiv:2107.08803* (2021).
- [130] Yen-Lun Liao, Xuanjun Chen, Chung-Che Wang, and Jyh-Shing Roger Jang. 2022. Adversarial speaker distillation for countermeasure model on Automatic speaker verification. *2nd Symposium on Security and Privacy in Speech Communication* (Sep 2022). <https://doi.org/10.21437/spsc.2022-6>
- [131] Suk-Young Lim, Dong-Kyu Chae, and Sang-Chul Lee. 2022. Detecting deepfake voice using explainable deep learning techniques. *Applied Sciences* 12, 8 (2022), 3926.
- [132] Haojian Lin, Yang Ai, and Zhenhua Ling. 2022. A Light CNN with Split Batch Normalization for Spoofed Speech Detection Using Data Augmentation. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 1684–1689.
- [133] Hanxiao Liu, Karen Simonyan, and Yiming Yang. 2018. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055* (2018).
- [134] Jie Liu, Zhibo Su, Hui Huang, Caiyan Wan, Quanxiu Wang, Jiangli Hong, Benlai Tang, and Fengjie Zhu. 2023. TranssionADD: A multi-frame reinforcement based sequence tagging model for audio deepfake detection. *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023)* (2023).
- [135] Rui Liu, Jinhua Zhang, Guanglai Gao, and Haizhou Li. 2023. Betray oneself: A novel audio deepfake detection model via mono-to-stereo conversion. *INTERSPEECH 2023* (Aug 2023). <https://doi.org/10.21437/interspeech.2023-2335>
- [136] Songxiang Liu, Haibin Wu, Hung-yi Lee, and Helen Meng. 2019. Adversarial attacks on spoofing countermeasures of automatic speaker verification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 312–319.
- [137] Wei Liu, Meng Sun, Xiongwei Zhang, Thomas Fang Zheng, et al. 2021. A multi-resolution front-end for end-to-end speech anti-spoofing. *arXiv preprint arXiv:2110.05087* (2021).
- [138] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems* 33 (2020), 21464–21475.
- [139] Xiaohui Liu, Meng Liu, Longbiao Wang, Kong Aik Lee, Hanyi Zhang, and Jianwu Dang. 2023. Leveraging positional-related local-global dependency for synthetic speech detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [140] Xiaohui Liu, Meng Liu, Lin Zhang, Linjuan Zhang, Chang Zeng, Kai Li, Nan Li, Kong Aik Lee, Longbiao Wang, and Jianwu Dang. 2022. Deep spectro-temporal artifacts for detecting synthesized speech. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. 69–75.
- [141] Xuechen Liu, Md Sahidullah, and Tomi Kinnunen. 2022. Spoofing-aware speaker verification with unsupervised domain adaptation. *The Speaker and Language Recognition Workshop (Odyssey 2022)* (Jun 2022). <https://doi.org/10.21437/odyssey.2022-12>
- [142] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, et al. 2023. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).
- [143] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling. 2018. The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods. In *The Speaker and Language Recognition Workshop*. ISCA, 195–202.
- [144] Jingze Lu, Zhuo Li, Yuxiang Zhang, Wenchao Wang, and Pengyuan Zhang. 2022. Acoustic or pattern? speech spoofing countermeasure based on image pre-training models. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. 77–84.
- [145] Jingze Lu, Yuxiang Zhang, Wenchao Wang, Zengqiang Shang, and Pengyuan Zhang. 2024. One-Class Knowledge Distillation for Spoofing Speech Detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 11251–11255.
- [146] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
- [147] Anwei Luo, Enlei Li, Yongliang Liu, Xiangui Kang, and Z. Jane Wang. 2021. A capsule network based approach for detection of audio spoofing attacks. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Jun 2021). <https://doi.org/10.1109/icassp39728.2021.9414670>
- [148] Zhiqiang Lv, Shanshan Zhang, Kai Tang, and Pengfei Hu. 2022. Fake audio detection based on unsupervised pretraining models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 9231–9235.
- [149] Zhiqiang Lv, Shanshan Zhang, Kai Tang, and Pengfei Hu. 2022. Fake audio detection based on unsupervised pretraining models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 9231–9235.

- [150] Haoxin Ma, Jiangyan Yi, Chenglong Wang, Xinrui Yan, Jianhua Tao, Tao Wang, Shiming Wang, Le Xu, and Ruibo Fu. 2022. FAD: A Chinese dataset for fake audio detection. *arXiv preprint arXiv:2207.12308* (2022).
- [151] Kaijie Ma, Yifan Feng, Beijing Chen, and Guoying Zhao. 2023. End-to-end dual-branch network towards synthetic speech detection. *IEEE Signal Processing Letters* 30 (2023), 359–363.
- [152] Qiaowei Ma, Jinghui Zhong, Yitao Yang, Weiheng Liu, Ying Gao, and Wing WY Ng. 2022. ConvNeXt Based Neural Network for Audio Anti-Spoofing. *arXiv preprint arXiv:2209.06434* (2022).
- [153] Xinyue Ma, Tianyu Liang, Shanshan Zhang, Shen Huang, and Liang He. 2021. Improved lightcnn with attention modules for ASV spoofing detection. *2021 IEEE International Conference on Multimedia and Expo (ICME)* (Jul 2021). <https://doi.org/10.1109/icme51207.2021.9428313>
- [154] Xinyue Ma, Shanshan Zhang, Shen Huang, Ji Gao, Ying Hu, and Liang He. 2023. How to boost anti-spoofing with X-vectors. *2022 IEEE Spoken Language Technology Workshop (SLT)* (Jan 2023). <https://doi.org/10.1109/slt54892.2023.10022504>
- [155] Youxuan Ma, Zongze Ren, and Shugong Xu. 2021. RW-resnet: A novel speech anti-spoofing model using Raw Waveform. *Interspeech 2021* (Aug 2021). <https://doi.org/10.21437/interspeech.2021-438>
- [156] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).
- [157] Daniele Mari, Federica Latora, and Simone Milani. 2022. The sound of silence: Efficiency of first digit features in synthetic audio detection. In *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 1–6.
- [158] Juan M. Martin-Donas and Aitor Alvarez. 2022. The Vicomtech audio deepfake detection system based on WAV2VEC2 for the 2022 add challenge. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (May 2022). <https://doi.org/10.1109/icassp43922.2022.9747768>
- [159] Juan Manuel Martín-Doñas and Aitor Álvarez. 2023. The Vicomtech partial deepfake detection and location system for the 2023 ADD Challenge. In *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*.
- [160] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. 2023. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence* 53, 4 (2023), 3974–4026.
- [161] Mvelo Mcuba, Avinash Singh, Richard Adeyemi Ikuesan, and Hein Venter. 2023. The effect of deep learning methods on deepfake audio detection for digital investigation. *Procedia Computer Science* 219 (2023), 211–219.
- [162] Fedila Meriem, Bengherabi Messaoud, and Yahya-zoubir Bahia. 2023. Texture analysis of edge mapped audio spectrogram for spoofing attack detection. *Multimedia Tools and Applications* (2023), 1–23.
- [163] Hiren Mewada, Jawad F Al-Asad, Faris A Almalki, Adil H Khan, Nouf Abdullah Almujally, Samir El-Nakla, and Qamar Naith. 2023. Gaussian-Filtered High-Frequency-Feature Trained Optimized BiLSTM Network for Spoofed-Speech Classification. *Sensors* 23, 14 (2023), 6637.
- [164] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition* 65 (2017), 211–222.
- [165] Nicolas M Müller, Pavel Czepin, Franziska Dieckmann, Adam Froggyar, and Konstantin Böttinger. 2022. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263* (2022).
- [166] Nicolas M Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. 2024. MLAAD: The Multi-Language Audio Anti-Spoofing Dataset. *arXiv preprint arXiv:2401.09512* (2024).
- [167] Sung Hwan Mun, Hye-jin Shim, Hemlata Tak, Xin Wang, Xuechen Liu, Md Sahidullah, Myeonghun Jeong, Min Hyun Han, Massimiliano Todisco, Kong Aik Lee, and et al. 2023. Towards single integrated spoofing-aware speaker Verification Embeddings. *INTERSPEECH 2023* (Aug 2023). <https://doi.org/10.21437/interspeech.2023-1402>
- [168] Arun Sankar Muttathu Sivasankara Pillai, Phillip L. De Leon, and Utz Roedig. 2022. Detection of voice conversion spoofing attacks using voiced speech. *Secure IT Systems* (2022), 159–175. https://doi.org/10.1007/978-3-031-22295-5_9
- [169] Nicolas Müller, Franziska Dieckmann, Pavel Czepin, Roman Canals, Konstantin Böttinger, and Jennifer Williams. 2021. Speech is silver, silence is golden: What do ASVSPOOF-trained models really learn? *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge* (Sep 2021). <https://doi.org/10.21437/asvspoof.2021-9>
- [170] Nicolas M. Müller, Philip Sperl, and Konstantin Böttinger. 2023. Complex-valued neural networks for voice anti-spoofing. In *Proc. INTERSPEECH 2023*. 3814–3818. <https://doi.org/10.21437/Interspeech.2023-901>
- [171] Long Nguyen-Vu, Thien-Phuc Doan, Mai Bui, Kihun Hong, and Souhwan Jung. 2023. On the defense of spoofing countermeasures against adversarial attacks. *IEEE Access* (2023).
- [172] Tijana Nosek, Siniša Suzić, Boris Papić, and Nikša Jakovljević. 2019. Synthesized speech detection based on spectrogram and convolutional neural networks. In *2019 27th Telecommunications Forum (TELFOR)*. IEEE, 1–4.
- [173] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. 2018. Attentive statistics pooling for deep speaker embedding. *arXiv preprint arXiv:1803.10963* (2018).

- [174] Monisankha Pal, Dipjyoti Paul, and Goutam Saha. 2018. Synthetic speech detection using fundamental frequency variation and spectral features. *Computer Speech & Language* 48 (2018), 31–50.
- [175] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5206–5210.
- [176] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. *Interspeech 2019* (Sep 2019). <https://doi.org/10.21437/interspeech.2019-2680>
- [177] Ankur T Patil, Hemant A Patil, and Kuldeep Khorja. 2022. Effectiveness of energy separation-based instantaneous frequency estimation for cochlear cepstral features for synthetic and voice-converted spoofed speech detection. *Computer Speech & Language* 72 (2022), 101301.
- [178] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
- [179] Michele Pilia, Sara Mandelli, Paolo Bestagini, and Stefano Tubaro. 2021. Time scaling detection and estimation in audio recordings. In *2021 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 1–6.
- [180] Md Hafizur Rahman, Martin Graciarena, Diego Castan, Chris Cobo-Kroenke, Mitchell McLaren, and Aaron Lawson. 2022. Detecting synthetic speech manipulation in real audio recordings. In *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 1–6.
- [181] Rishabh Ranjan, Mayank Vatsa, and Richa Singh. 2022. Statnet: Spectral and temporal features based multi-task network for audio spoofing detection. *2022 IEEE International Joint Conference on Biometrics (IJCB)* (Oct 2022). <https://doi.org/10.1109/ijcb54206.2022.10007949>
- [182] Ruchira Ray, Sanka Karthik, Vinayak Mathur, Prashant Kumar, G Maragatham, Sourabh Tiwari, and Rashmi T Shankarappa. 2021. Feature genuinization based residual squeeze-and-excitation for audio anti-spoofing in sound AI. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 1–5.
- [183] Ricardo Reimao and Vassilios Tzerpos. 2019. For: A dataset for synthetic speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 1–10.
- [184] Yeqing Ren, Haipeng Peng, Lixiang Li, Xiaopeng Xue, Yang Lan, and Yixian Yang. 2023. A voice spoofing detection framework for IoT systems with feature pyramid and online knowledge distillation. *Journal of Systems Architecture* 143 (2023), 102981.
- [185] Yeqing Ren, Haipeng Peng, Lixiang Li, and Yixian Yang. 2023. Lightweight Voice Spoofing Detection using Improved One-Class Learning and Knowledge Distillation. *IEEE Transactions on Multimedia* (2023).
- [186] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III* 18. Springer, 234–241.
- [187] Amir Mohammad Rostami, Mohammad Mehdi Homayounpour, and Ahmad Nickabadi. 2023. Efficient attention branch network with combined loss function for automatic speaker verification spoof detection. *Circuits, Systems, and Signal Processing* 42, 7 (2023), 4252–4270.
- [188] Md. Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. 2015. A comparison of features for synthetic speech detection. In *Proc. Interspeech 2015*. 2087–2091. <https://doi.org/10.21437/Interspeech.2015-472>
- [189] Davide Salvi, Paolo Bestagini, and Stefano Tubaro. 2023. Reliability Estimation for Synthetic Speech Detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [190] Davide Salvi, Brian Hosler, Paolo Bestagini, Matthew C Stamm, and Stefano Tubaro. 2023. TIMIT-TTS: a Text-to-Speech Dataset for Multimodal Synthetic Media Detection. *IEEE Access* (2023).
- [191] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
- [192] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2021. AISHELL-3: A Multi-Speaker Mandarin TTS Corpus. In *Proc. Interspeech 2021*. 2756–2760. <https://doi.org/10.21437/Interspeech.2021-755>
- [193] Hye-jin Shim, Jungwoo Heo, Jae-Han Park, Ga-Hui Lee, and Ha-Jin Yu. 2022. Graph attentive feature aggregation for text-independent speaker verification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7972–7976.
- [194] Hye-jin Shim, Jee-weon Jung, and Tomi Kinnunen. 2023. Multi-dataset co-training with sharpness-aware optimization for audio anti-spoofing. *INTERSPEECH 2023* (Aug 2023). <https://doi.org/10.21437/interspeech.2023-1910>
- [195] Hye-jin Shim, Hemlata Tak, Xuechen Liu, Hee-Soo Heo, Jee-weon Jung, Joon Son Chung, Soo-Whan Chung, Ha-Jin Yu, Bong-Jin Lee, Massimiliano Todisco, and et al. 2022. Baseline Systems for the first spoofing-aware speaker Verification Challenge: Score and Embedding Fusion. *The Speaker and Language Recognition Workshop (Odyssey 2022)*

- (Jun 2022). <https://doi.org/10.21437/odyssey.2022-46>
- [196] Arun Kumar Singh and Priyanka Singh. 2021. Detection of ai-synthesized speech using cepstral & bispectral statistics. In *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 412–417.
 - [197] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*. PMLR, 4693–4702.
 - [198] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. 2017. Deep Neural Network Embeddings for Text-Independent Speaker Verification. In *Proc. Interspeech 2017*. 999–1003. <https://doi.org/10.21437/Interspeech.2017-620>
 - [199] Meet H. Soni, Tanvina B. Patel, and Hemant A. Patil. 2016. Novel Subband Autoencoder features for detection of spoofed speech. *Interspeech 2016* (Sep 2016). <https://doi.org/10.21437/interspeech.2016-668>
 - [200] Kaavya Sriskandaraja, Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Haizhou Li. 2016. Front-end for anti-spoofing countermeasures in speaker verification: Scattering spectral decomposition. *IEEE Journal of Selected Topics in Signal Processing* 11, 4 (2016), 632–643.
 - [201] Kaavya Sriskandaraja, Vidhyasaharan Sethu, Phu Ngoc Le, and Eliathamby Ambikairajah. 2016. Investigation of sub-band discriminative information between spoofed and genuine speech. *Interspeech 2016* (Sep 2016). <https://doi.org/10.21437/interspeech.2016-844>
 - [202] Nishant Subramani and Delip Rao. 2020. Learning efficient representations for fake speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5859–5866.
 - [203] Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.
 - [204] Hemlata Tak, Jee-weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans. 2021. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. *arXiv preprint arXiv:2107.12710* (2021).
 - [205] Hemlata Tak, Jee-weon Jung, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2021. Graph attention networks for anti-spoofing. *Interspeech 2021* (Aug 2021). <https://doi.org/10.21437/interspeech.2021-993>
 - [206] Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2022. Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6382–6386.
 - [207] Hemlata Tak, Jose Patino, Andreas Nautsch, Nicholas W. D. Evans, and Massimiliano Todisco. 2020. An Explainability Study of the Constant Q Cepstral Coefficient Spoofing Countermeasure for Automatic Speaker Verification. In *Odyssey 2020: The Speaker and Language Recognition Workshop, 1-5 November 2020, Tokyo, Japan*, Kong-Aik Lee, Takafumi Koshinaka, and Koichi Shinoda (Eds.). ISCA, 333–340. <https://doi.org/10.21437/Odyssey.2020-47>
 - [208] 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.
 - [209] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. 2022. Automatic speaker verification spoofing and Deepfake detection using WAV2VEC 2.0 and data augmentation. *The Speaker and Language Recognition Workshop (Odyssey 2022)* (Jun 2022). <https://doi.org/10.21437/odyssey.2022-16>
 - [210] Pablo Andrés Tamayo Flórez et al. 2022. Voice anti-spoofing data-set built from Latin American Spanish accents implementing voice conversion and text-to-speech techniques. (2022).
 - [211] Zhongwei Teng, Quchen Fu, Jules White, Maria Powell, and Douglas Schmidt. 2022. Sa-SASV: An end-to-end spoof-aggregated spoofing-aware speaker verification system. *Interspeech 2022* (Sep 2022). <https://doi.org/10.21437/interspeech.2022-11029>
 - [212] 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.
 - [213] Anton Tomilov, Aleksei Svishchev, Marina Volkova, Artem Chirkovskiy, Alexander Kondratev, and Galina Lavrentyeva. 2021. STC antispoofing systems for the ASVSPOOF2021 challenge. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge* (Sep 2021). <https://doi.org/10.21437/asvspoof.2021-10>
 - [214] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* 12 (2016).
 - [215] Marina Volkova, Tseren Andzhukaev, Galina Lavrentyeva, Sergey Novoselov, and Alexander Kozlov. 2019. Light CNN architecture enhancement for different types spoofing attack detection. In *Speech and Computer: 21st International Conference, SPECOM 2019, Istanbul, Turkey, August 20–25, 2019, Proceedings 21*. Springer, 520–529.

- [216] Chenglong Wang, Jiayi He, Jiangyan Yi, Jianhua Tao, Chu Yuan Zhang, and Xiaohui Zhang. 2024. Multi-scale permutation entropy for audio deepfake detection. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Apr 2024). <https://doi.org/10.1109/icassp48485.2024.10448095>
- [217] Chenglong Wang, Jiangyan Yi, Jianhua Tao, Haiyang Sun, Xun Chen, Zhengkun Tian, Haoxin Ma, Cunhang Fan, and Ruibo Fu. 2022. Fully automated end-to-end fake audio detection. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. 27–33.
- [218] Chenglong Wang, Jiangyan Yi, Jianhua Tao, Chu Yuan Zhang, Shuai Zhang, and Xun Chen. 2023. Detection of cross-dataset fake audio based on prosodic and pronunciation features. *INTERSPEECH 2023* (Aug 2023). <https://doi.org/10.21437/interspeech.2023-1254>
- [219] Chenglong Wang, Jiangyan Yi, Xiaohui Zhang, Jianhua Tao, Le Xu, and Ruibo Fu. 2023. Low-rank adaptation method for wav2vec2-based fake audio detection. *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023)* (2023).
- [220] Lei Wang, Benedict Yeoh, and Jun Wah Ng. 2022. Synthetic voice detection and audio splicing detection using se-res2net-conformer architecture. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 115–119.
- [221] Longbiao Wang, Yohei Yoshida, Yuta Kawakami, and Seiichi Nakagawa. 2015. Relative phase information for detecting human speech and spoofed speech.. In *INTERSPEECH*. 2092–2096.
- [222] Ruoyu Wang, Jun Du, and Tian Gao. 2023. Quantum transfer learning using the large-scale unsupervised pre-trained model wavlm-large for synthetic speech detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [223] Ruoyu Wang, Jun Du, and Chang Wang. 2022. Multi-branch Network with Circle Loss Using Voice Conversion and Channel Robust Data Augmentation for Synthetic Speech Detection. In *Chinese Conference on Biometric Recognition*. Springer, 613–620.
- [224] Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, and Yang Liu. 2020. Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. In *Proceedings of the 28th ACM international conference on multimedia*. 1207–1216.
- [225] Xingming Wang, Xiaoyi Qin, Yikang Wang, Yunfei Xu, and Ming Li. 2022. The DKU-oppo system for the 2022 spoofing-aware speaker Verification Challenge. *Interspeech 2022* (Sep 2022). <https://doi.org/10.21437/interspeech.2022-11190>
- [226] Xin Wang and Junichi Yamagishi. 2021. A comparative study on recent neural spoofing countermeasures for synthetic speech detection. *Interspeech 2021* (Aug 2021). <https://doi.org/10.21437/interspeech.2021-702>
- [227] Xin Wang and Junichi Yamagishi. 2022. Investigating self-supervised front ends for speech spoofing countermeasures. *The Speaker and Language Recognition Workshop (Odyssey 2022)* (Jun 2022). <https://doi.org/10.21437/odyssey.2022-14>
- [228] Xin Wang and Junichi Yamagishi. 2022. A practical guide to logical access voice presentation attack detection. In *Frontiers in Fake Media Generation and Detection*. Springer, 169–214.
- [229] Xin Wang and Junichi Yamagishi. 2023. Investigating active-learning-based training data selection for speech spoofing countermeasure. In *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 585–592.
- [230] Xin Wang and Junichi Yamagishi. 2023. Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Jun 2023). <https://doi.org/10.1109/icassp49357.2023.10094779>
- [231] Xin Wang and Junichi Yamagishi. 2024. Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end? *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Apr 2024). <https://doi.org/10.1109/icassp48485.2024.10446331>
- [232] Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards End-to-End Speech Synthesis. In *Proc. Interspeech 2017*. 4006–4010. <https://doi.org/10.21437/Interspeech.2017-1452>
- [233] Yikang Wang, Xingming Wang, Hiromitsu Nishizaki, and Ming Li. 2022. Low pass filtering and bandwidth extension for robust anti-spoofing countermeasure against codec variabilities. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 438–442.
- [234] 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.
- [235] Ziqian Wang, Qing Wang, Jixun Yao, and Lei Xie. 2022. The NPU-ASLP System for Deepfake Algorithm Recognition in ADD 2023 Challenge. *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023)* (2022).
- [236] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.

- [237] Haibin Wu, Jiawen Kang, Lingwei Meng, Helen Meng, and Hung-yi Lee. 2023. The defender's perspective on automatic speaker verification: An overview. *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023)* (2023).
- [238] Haibin Wu, Jiawen Kang, Lingwei Meng, Yang Zhang, Xixin Wu, Zhiyong Wu, Hung-yi Lee, and Helen Meng. 2022. Tackling spoofing-aware speaker verification with multi-model fusion. *The Speaker and Language Recognition Workshop (Odyssey 2022)* (Jun 2022). <https://doi.org/10.21437/odyssey.2022-13>
- [239] Haibin Wu, Heng-Cheng Kuo, Naijun Zheng, Kuo-Hsuan Hung, Hung-Yi Lee, Yu Tsao, Hsin-Min Wang, and Helen Meng. 2022. Partially fake audio detection by self-attention-based fake span discovery. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 9236–9240.
- [240] Haochen Wu, Zhuohai Li, Luzhen Xu, Zhentao Zhang, Wenting Zhao, Bin Gu, Yang Ai, Yexin Lu, Jie Zhang, Zhenhua Ling, et al. 2022. The USTC-NERC SLIP System for the Track 1.2 of Audio Deepfake Detection (ADD 2023) Challenge. *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023)* (2022).
- [241] Haibin Wu, Songxiang Liu, Helen Meng, and Hung-yi Lee. 2020. Defense against adversarial attacks on spoofing countermeasures of asv. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6564–6568.
- [242] Haibin Wu, Lingwei Meng, Jiawen Kang, Jinchao Li, Xu Li, Xixin Wu, Hung-yi Lee, and Helen Meng. 2022. Spoofing-aware speaker verification by multi-level fusion. *Interspeech 2022* (Sep 2022). <https://doi.org/10.21437/interspeech.2022-920>
- [243] Lei Wu and Ye Jiang. 2022. Attentional Fusion TDNN for Spoof Speech Detection. In *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*. IEEE, 651–657.
- [244] Zhenzong Wu, Rohan Kumar Das, Jichen Yang, and Haizhou Li. 2020. Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. *Interspeech 2020* (Oct 2020). <https://doi.org/10.21437/interspeech.2020-1810>
- [245] Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li. 2013. Synthetic speech detection using temporal modulation feature. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 7234–7238. <https://doi.org/10.1109/ICASSP.2013.6639067>
- [246] Ziyue Xiang, Amit Kumar Singh Yadav, Stefano Tubaro, Paolo Bestagini, and Edward J Delp. 2023. Extracting efficient spectrograms from MP3 compressed speech signals for synthetic speech detection. In *Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security*. 163–168.
- [247] Yuankun Xie, Haonan Cheng, Yutian Wang, and Long Ye. 2022. Single domain generalization for audio deepfake detection. *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023)* (2022).
- [248] Yuankun Xie, Haonan Cheng, Yutian Wang, and Long Ye. 2023. Learning a self-supervised domain-invariant feature representation for generalized audio deepfake detection. *INTERSPEECH 2023* (Aug 2023). <https://doi.org/10.21437/interspeech.2023-1383>
- [249] Yang Xie, Zhenchuan Zhang, and Yingchun Yang. 2021. Siamese network with wav2vec feature for spoofing speech detection. *Interspeech 2021* (Aug 2021). <https://doi.org/10.21437/interspeech.2021-847>
- [250] Jun Xue, Cunhang Fan, Zhao Lv, Jianhua Tao, Jiangyan Yi, Chengshi Zheng, Zhengqi Wen, Minmin Yuan, and Shengang Shao. 2022. Audio deepfake detection based on a combination of f0 information and real plus imaginary spectrogram features. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. 19–26.
- [251] Jun Xue, Cunhang Fan, Jiangyan Yi, Chenglong Wang, Zhengqi Wen, Dan Zhang, and Zhao Lv. 2023. Learning from yourself: A self-distillation method for fake speech detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [252] Junxiao Xue, Hao Zhou, Huawei Song, Bin Wu, and Lei Shi. 2023. Cross-modal information fusion for voice spoofing detection. *Speech Communication* 147 (2023), 41–50.
- [253] Amit Kumar Singh Yadav, Emily R Bartusiak, Kratika Bhagtani, and Edward J Delp. 2023. Synthetic speech attribution using self supervised audio spectrogram transformer. *Electronic Imaging* 35 (2023), 1–11.
- [254] Amit Kumar Singh Yadav, Ziyue Xiang, Emily R Bartusiak, Paolo Bestagini, Stefano Tubaro, and Edward J Delp. 2023. ASSD: Synthetic Speech Detection in the AAC Compressed Domain. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [255] Junichi Yamagishi, Massimiliano Todisco, Md Sahidullah, Héctor Delgado, Xin Wang, Nicolas Evans, Tomi Kinnunen, Kong Aik Lee, Ville Vestman, and Andreas Nautsch. 2019. Asvspoof 2019: The 3rd automatic speaker verification spoofing and countermeasures challenge database. (2019).
- [256] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *University of Edinburgh. The Centre for Speech Technology Research (CSTR)* (2019).
- [257] Rui Yan, Cheng Wen, Shuran Zhou, Tingwei Guo, Wei Zou, and Xiangang Li. 2022. Audio deepfake detection system with neural stitching for add 2022. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal*

- Processing (ICASSP)*. IEEE, 9226–9230.
- [258] Jichen Yang and Rohan Kumar Das. 2020. Long-term high frequency features for synthetic speech detection. *Digital Signal Processing* 97 (2020), 102622.
 - [259] Jichen Yang, Rohan Kumar Das, and Haizhou Li. 2018. Extended constant-Q cepstral coefficients for detection of spoofing attacks. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 1024–1029.
 - [260] Jichen Yang, Rohan Kumar Das, and Haizhou Li. 2019. Significance of subband features for synthetic speech detection. *IEEE Transactions on Information Forensics and Security* 15 (2019), 2160–2170.
 - [261] Jichen Yang, Hongji Wang, Rohan Kumar Das, and Yanmin Qian. 2021. Modified magnitude-phase spectrum information for spoofing detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1065–1078.
 - [262] Minjiao Yang, Kangfeng Zheng, Xiujuan Wang, Yudao Sun, and Zhe Chen. 2023. Comparative analysis of asv spoofing countermeasures: Evaluating res2net-based approaches. *IEEE Signal Processing Letters* (2023).
 - [263] Yujie Yang, Haochen Qin, Hang Zhou, Chengcheng Wang, Tianyu Guo, Kai Han, and Yunhe Wang. 2024. A robust audio deepfake detection system via multi-view feature. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Apr 2024). <https://doi.org/10.1109/icassp48485.2024.10446560>
 - [264] Jiangyan Yi, Ye Bai, Jianhua Tao, Zhengkun Tian, Chenglong Wang, Tao Wang, and Ruibo Fu. 2021. Half-truth: A partially fake audio detection dataset. *arXiv preprint arXiv:2104.03617* (2021).
 - [265] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al. 2022. Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 9216–9220.
 - [266] Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, et al. 2023. ADD 2023: the Second Audio Deepfake Detection Challenge. *arXiv preprint arXiv:2305.13774* (2023).
 - [267] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. 2023. Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970* (2023).
 - [268] Zhao Yi, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhenhua Ling, and Tomoki Toda. 2020. Voice conversion challenge 2020—intra-lingual semi-parallel and cross-lingual voice conversion—. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, Vol. 2020. 80–98.
 - [269] Hong Yu, Achintya Sarkar, Dennis Alexander Lehmann Thomsen, Zheng-Hua Tan, Zhanyu Ma, and Jun Guo. 2016. Effect of multi-condition training and speech enhancement methods on spoofing detection. In *2016 First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE)*. IEEE, 1–5.
 - [270] Hong Yu, Zheng-Hua Tan, Yiming Zhang, Zhanyu Ma, and Jun Guo. 2017. DNN filter bank cepstral coefficients for spoofing detection. *Ieee Access* 5 (2017), 4779–4787.
 - [271] Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi. 2021. LEAF: A learnable frontend for audio classification. *arXiv preprint arXiv:2101.08596* (2021).
 - [272] Neil Zeghidour, Nicolas Usunier, Iasonas Kokkinos, Thomas Schaiz, Gabriel Synnaeve, and Emmanuel Dupoux. 2018. Learning filterbanks from raw speech for phone recognition. In *2018 IEEE international conference on acoustics, speech and signal Processing (ICASSP)*. IEEE, 5509–5513.
 - [273] Hossein Zeinali, Themis Stafylakis, Georgia Athanasopoulou, Johan Rohdin, Ioannis Gkinis, Lukáš Burget, and Jan Černocký. 2019. Detecting spoofing attacks using VGG and SincNet: But-omilia submission to ASVSPOOF 2019 challenge. *Interspeech 2019* (Sep 2019). <https://doi.org/10.21437/interspeech.2019-2892>
 - [274] Xiao-Min Zeng, Jiang-Tao Zhang, Kang Li, Zhuo-Li Liu, Wei-Lin Xie, and Yan Song. 2023. Deepfake Algorithm Recognition System with Augmented Data for ADD 2023 Challenge. In *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*.
 - [275] Bowen Zhang and Terence Sim. 2022. Localizing fake segments in speech. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 3224–3230.
 - [276] Jiachen Zhang, Guoqing Tu, Shubo Liu, and Zhaoxue Cai. 2023. Audio Anti-Spoofing Based on Audio Feature Fusion. *Algorithms* 16, 7 (2023), 317.
 - [277] Li Zhang, Yue Li, Huan Zhao, Qing Wang, and Lei Xie. 2022. Backend ensemble for speaker verification and spoofing countermeasure. *Interspeech 2022* (Sep 2022). <https://doi.org/10.21437/interspeech.2022-10259>
 - [278] Lin Zhang, Xin Wang, Erica Cooper, Nicholas Evans, and Junichi Yamagishi. 2022. The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2022), 813–825.
 - [279] Lin Zhang, Xin Wang, Erica Cooper, Nicholas Evans, and Junichi Yamagishi. 2023. Range-Based Equal Error Rate for Spoof Localization. In *Proc. INTERSPEECH 2023*. 3212–3216. <https://doi.org/10.21437/Interspeech.2023-1214>

- [280] Lin Zhang, Xin Wang, Erica Cooper, and Junichi Yamagishi. 2021. Multi-task learning in utterance-level and segmental-level spoof detection. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge* (Sep 2021). <https://doi.org/10.21437/asvspoof.2021-2>
- [281] Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi, Jose Patino, and Nicholas Evans. 2021. An initial investigation for detecting partially spoofed audio. *Interspeech 2021* (Aug 2021). <https://doi.org/10.21437/interspeech.2021-738>
- [282] Peng Zhang, Peng Hu, and Xueliang Zhang. 2022. Norm-constrained score-level ensemble for spoofing aware speaker verification. *Interspeech 2022* (Sep 2022). <https://doi.org/10.21437/interspeech.2022-470>
- [283] Tao Zhang. 2022. Deepfake generation and detection, a survey. *Multimedia Tools and Applications* 81, 5 (2022), 6259–6276.
- [284] Teng Zhang, Lirui Deng, Liang Zhang, and Xianglei Dang. 2020. Deep learning in face synthesis: A survey on deepfakes. In *2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology (CCET)*. IEEE, 67–70.
- [285] Xiaohui Zhang, Jiangyan Yi, Jianhua Tao, Chenlong Wang, Le Xu, and Ruibo Fu. 2023. Adaptive Fake Audio Detection with Low-Rank Model Squeezing. *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023)* (2023).
- [286] Xiaohui Zhang, Jiangyan Yi, Jianhua Tao, Chenglong Wang, and Chu Yuan Zhang. 2023. Do you remember? Overcoming catastrophic forgetting for fake audio detection. In *International Conference on Machine Learning*. PMLR, 41819–41831.
- [287] You Zhang, Fei Jiang, and Zhiyao Duan. 2021. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters* 28 (2021), 937–941. <https://doi.org/10.1109/lsp.2021.3076358>
- [288] Yuxiang Zhang, Zhuo Li, Jingze Lu, Hua Hua, Wenchao Wang, and Pengyuan Zhang. 2023. The Impact of Silence on Speech Anti-Spoofing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).
- [289] Yuxiang Zhang, Zhuo Li, Jingze Lu, Wenchao Wang, and Pengyuan Zhang. 2024. Synthetic speech detection based on the temporal consistency of speaker features. *IEEE Signal Processing Letters* 31 (2024), 944–948. <https://doi.org/10.1109/lsp.2024.3381890>
- [290] Yuxiang Zhang, Zhuo Li, Wenchao Wang, and Pengyuan Zhang. 2022. SASV based on pre-trained ASV system and Integrated Scoring Module. *Interspeech 2022* (Sep 2022). <https://doi.org/10.21437/interspeech.2022-10149>
- [291] Yuxiang Zhang, Wenchao Wang, and Pengyuan Zhang. 2021. The effect of silence and dual-band fusion in anti-spoofing system. *Interspeech 2021* (Aug 2021). <https://doi.org/10.21437/interspeech.2021-1281>
- [292] You Zhang, Ge Zhu, and Zhiyao Duan. 2022. A probabilistic fusion framework for spoofing aware speaker verification. *The Speaker and Language Recognition Workshop (Odyssey 2022)* (Jun 2022). <https://doi.org/10.21437/odyssey.2022-11>
- [293] You Zhang, Ge Zhu, Fei Jiang, and Zhiyao Duan. 2021. An empirical study on channel effects for synthetic voice spoofing countermeasure systems. *Interspeech 2021* (Aug 2021). <https://doi.org/10.21437/interspeech.2021-1820>
- [294] Zhenyu Zhang, Yewei Gu, Xiaowei Yi, and Xianfeng Zhao. 2021. FMFCC-a: a challenging Mandarin dataset for synthetic speech detection. In *International Workshop on Digital Watermarking*. Springer, 117–131.
- [295] Zhenyu Zhang, Xiaowei Yi, and Xianfeng Zhao. 2021. Fake speech detection using residual network with transformer encoder. In *Proceedings of the 2021 ACM workshop on information hiding and multimedia security*. 13–22.
- [296] Zhenyu Zhang, Xianfeng Zhao, and Xiaowei Yi. 2022. Improving robustness of speech anti-spoofing system using resnext with neighbor filters. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.
- [297] Ye Zhou, Jianwu Zhang, and Pengguo Zhang. 2022. Spoof speech detection based on raw cross-dimension interaction attention network. In *Chinese Conference on Biometric Recognition*. Springer, 621–629.
- [298] Yi Zhu, Saurabh Powar, and Tiago H Falk. 2023. Characterizing the temporal dynamics of universal speech representations for generalizable deepfake detection. *arXiv preprint arXiv:2309.08099* (2023).
- [299] Yupeng Zhu, Zuxing Zhao, Fan Li, and Yanxiang Chen. 2023. Unseen Codec Spoof Speech Detection Based on Channel-Robust Feature. In *Proceedings of the 2023 3rd International Conference on Artificial Intelligence, Automation and Algorithms*. 10–14.