Computer Science Review 57 (2025) 100757

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Review article

A comprehensive survey with critical analysis for deepfake speech detection

Lam Pham[a],[1], Phat Lam[b],[1], Dat Tran[c],[1], Hieu Tang[d], Tin Nguyen[b], Alexander Schindler[a], Florian Skopik[a], Alexander Polonsky[e], Hai Canh Vu[f],*

[a] Austrian Institute of Technology (AIT), Vienna, Austria
[b] Ho Chi Minh City University of Technology, Ho Chi Minh City, Viet Nam
[c] FPT University, Ho Chi Minh City, Viet Nam
[d] University of Technology of Troyes, Aube, France
[e] BLOOM Social Analytics, France
[f] Laboratoire Roberval, Université de technologie de Compiègne, France

ARTICLE INFO

Keywords:
Deepfake speech detection (DSD)
Challenge competition
Ensemble
Audio embedding
Pre-trained model

ABSTRACT

Thanks to advancements in deep learning, speech generation systems now power a variety of real-world applications, such as text-to-speech for individuals with speech disorders, voice chatbots in call centers, cross-linguistic speech translation, etc. While these systems can autonomously generate human-like speech and replicate specific voices, they also pose risks when misused for malicious purposes. This motivates the research community to develop models for detecting synthesized speech (e.g., fake speech) generated by deep-learning-based models, referred to as the Deepfake Speech Detection task. As the Deepfake Speech Detection task has emerged in recent years, there are not many survey papers proposed for this task. Additionally, existing surveys for the Deepfake Speech Detection task tend to summarize techniques used to construct a Deepfake Speech Detection system rather than providing a thorough analysis. This gap motivated us to conduct a comprehensive survey, providing a critical analysis of the challenges and developments in Deepfake Speech Detection (This work is a part of our projects of STARLIGHT, EUCINF, and DEFAME FAKEs). Our survey is innovatively structured, offering an in-depth analysis of current challenge competitions, public datasets, and the deep-learning techniques that provide enhanced solutions to address existing challenges in the field. From our analysis, we propose hypotheses on leveraging and combining specific deep learning techniques to improve the effectiveness of Deepfake Speech Detection systems. Beyond conducting a survey, we perform extensive experiments to validate these hypotheses and propose a highly competitive model for the task of Deepfake Speech Detection. Given the analysis and the experimental results, we finally indicate potential and promising research directions for the Deepfake Speech Detection task.

Contents

* Corresponding author.
E-mail address: hai-canh.vu@utc.fr (H.C. Vu).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.cosrev.2025.100757
Received 28 December 2024; Received in revised form 17 March 2025; Accepted 23 March 2025
Available online 16 April 2025

1. Introduction

In recent years, remarkable advancements in deep learning techniques and neural networks have revolutionized the field of generative AI. Today, core communication mediums such as audio, images, video, and text can be automatically generated and applied across various domains, including chatbot systems (e.g., ChatGPT), film production [10], code generation [11], and audio synthesis [12,13], etc. However, AI-synthesized data could pose a serious threat to social security when there is an increasing number of crimes related to leveraging the synthesized data [14]. To address this concern, the tasks, which are proposed for detecting synthesized data (e.g. fake data) generated from deep-learning-based methods, referred to as deepfake detection, have drawn much attention from the research community recently.

Focusing on human speech, this paper provides a comprehensive survey for the task of Deepfake Speech Detection (DSD). To this end, the milestones presenting the development progress of the DSD task are first presented in Fig. 1. As the figure shows, the earliest public dataset and challenge proposed for the DSD task was introduced in 2015, focusing exclusively on the English language. Then, the first challenge for video deepfake detection (DFDC) [15] was introduced in 2020. In subsequent years, datasets for the DSD task in Japanese [16], Korean [16], and Chinese [17] were introduced in 2021 and 2022, respectively. Recently, in 2024, multilingual datasets for the DSD task have been published, including MLAAD [18] for conversational speech and SVDD [19] for singing. Fig. 1 also highlights a growing number of papers, datasets, and challenge competitions for the DSD task from 2021 to the present. This trend indicates that the DSD task has recently gained prominence and has attracted significant interest from the research community. To further understand the DSD task, we summarized recent survey papers related to the DSD task in Table 1. As shown in the table, most of these surveys focus on detecting general fake data (e.g., images, videos, audio, or text), with audio or human speech typically being addressed only as a subsection or a part of the broader discussion [2,3,8]. Therefore, the main techniques, existing concerns, and potential research for the DSD task have not been comprehensively analyzed in these papers. Among the survey papers, only two survey papers of [5,9] focus on the DSD task. However, as conventional surveys, these papers primarily summarize the technologies used to construct a DSD system such as datasets, feature extraction, classification model, loss functions, rather than providing a comprehensive analysis and highlighting existing concerns. For instance, while challenge competitions proposed for the DSD task are very important in advancing the research community, their importance and various aspects have not been thoroughly analyzed (e.g., the number of research teams participating in these competitions and their results are interesting to analyze. Although this information reflects the level of interest in DSD within the research community, it has not been addressed in any existing survey papers). The second concern is related to public datasets proposed for the DSD task. In particular, the current survey papers do not adequately analyze the imbalance among (1) the number of utterances, (2) the AI-synthesized speech systems used to generate fake speech, and (3) the original/real human speech resource used to generate fake speech utterances. These key factors are essential in creating a high-quality DSD dataset for evaluating DSD models. Additionally, survey papers are at risk of becoming outdated as new datasets, techniques, and models continue to emerge. However, current surveys do not offer solutions for regularly updating essential information, such as details about challenge competitions, public datasets, and the top-performing models on specific datasets. Regarding technologies used to construct a DSD model such as feature extraction, classification model, or loss functions, current survey papers mainly summarize and then present conclusions rather than conducting experiments to provide strong evidence and validation.

The above concerns about the existing survey papers for the DSD task motivate and inspire us to provide a much more comprehensive survey in this paper. By addressing these concerns, we make the following main contributions:

- We provide a comprehensive analysis and then indicate concerns related to three main topics: The current challenge competition,

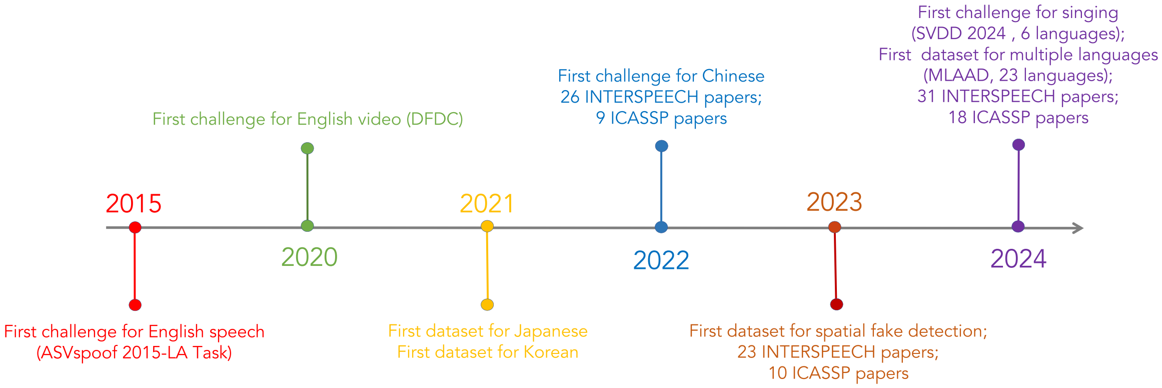
First challenge for singing
(SVDD 2024 , 6 languages);
First dataset for multiple languages
(MLAAD, 23 languages);
31 INTERSPEECH papers;
18 ICASSP papers
First challenge for Chinese
26 INTERSPEECH papers;
9 ICASSP papers
First challenge for English video (DFDC)
2015
2021
2023
2020
2022
2024
First challenge for English speech
(ASVspoof 2015-LA Task)
First dataset for Japanese
First dataset for Korean
First dataset for spatial fake detection;
23 INTERSPEECH papers;
10 ICASSP papers

Fig. 1. The timeline of Deepfake Speech Detection (DSD) task.

Table 1
The main factors analyzed in survey papers.

Papers	Years	Audio/Video	Challenge competitions	Public datasets	Data augmentation	Feature extraction	Classification models	Loss functions	Training strategies	Proposed models	Continue updating
[1]	2021	Yes/Yes	No	Yes	No	No	Yes	No	No	No	No
[2]	2023	Yes/Yes	No	No	No	Yes	Yes	No	No	No	No
[3]	2023	Yes/Yes	No	No	No	Yes	Yes	Yes	No	No	No
[4]	2023	Yes/Yes	No	Yes	No	No	Yes	Yes	Yes	No	No
[5]	2023	Yes/No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No
[6]	2023	Yes/Yes	No	Yes	No	No	Yes	No	No	No	No
[7]	2024	Yes/Yes	No	Yes	No	No	Yes	No	No	No	No
[8]	2024	Yes/Yes	No	Yes	No	Yes	Yes	No	No	No	No
[9]	2024	Yes/No	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Our survey	**2024**	**Yes/No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

the published datasets, and the deep-learning-based techniques used to develop a DSD system. Each topic consists of three main parts: ‘Analysis’, ‘Discussion’, and ‘Conclusion’. The ‘Analysis’ summarizes concrete information about the topic. The ‘Discussion’ indicates concerns in each topic. Finally, the ‘Conclusion’ provides a summary of what we discussed and indicates some insights to further improve each topic.

- To solve the out-of-date issue of a survey paper, we set up a Github repository to update further challenge competitions, public datasets, and top-performance systems. New versions of the paper are also continually updated on ‘https://arxiv.org’.
- More than a survey, we conduct extensive experiments to verify assumptions from the comprehensive analysis (i.e., different types of data augmentation, multiple input features, multiple network architectures, cross-dataset and cross-language evaluation, etc.), achieving a competitive DSD model. Given the analysis and experimental results, we indicate potential research directions for the DSD task.

The remainder of this paper is structured as follows: Section 2 discusses challenge competitions for the DSD task. Section 3 deeply analyses the public and benchmark datasets proposed for the DSD task. In Section 4, we summarize the key techniques for constructing the main components of a DSD system, including data augmentation, feature extraction, classification models, and loss functions Section 5 presents extensive experiments that validate the techniques described in Section 4. Building on the analysis and results from the previous sections, Section 6 outlines our proposed research directions in the DSD task. Finally, Section 7 concludes the paper.

2. Challenge competitions proposed for deepfake speech detection

Analysis: Challenge competitions for the DSD task play a crucial role in motivating the research community. These competitions not only introduce new benchmark datasets but also host workshops where research teams can discuss their ideas and share their motivations. This environment encourages the community to publish more datasets and develop new techniques to address the DSD challenges. To analyze DSD challenge competitions, we first summarize all challenges in Table 2. Importantly, we will continually update information about future DSD challenge competitions in our GitHub repository.[2]

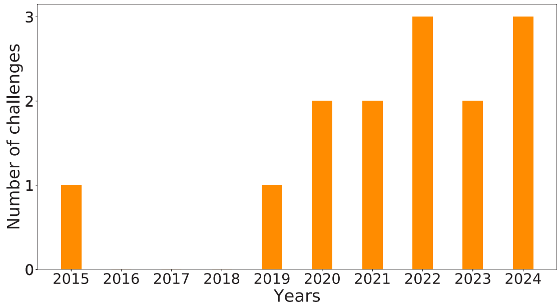

Fig. 2. The number of competitions proposed for DSD task from 2015.

As Table 2 shows, most challenge competitions focus on detecting fake speech in a conversation except for the SVDD 2024 challenge [28] for the fake singing detection. All challenge competitions for fake speech detection in a conversation have been proposed for a single language (i.e., While ADD 2022 and ADD 2023 are for Chinese, the others are proposed for English). Regarding the number of DSD challenge competitions, Fig. 2 shows that there has been an increase in recent years. This trend indicates that the DSD task has gained attention from the research community, particularly due to the rise of advanced deep learning systems capable of generating highly realistic human-like speech, which poses significant security risks. DSD

[2] https://github.com/AI-ResearchGroup/A-Comprehensive-Survey-with-Critical-Analysis-for-Deepfake-Speech-Detection

Table 2
The challenge competitions proposed for Deepfake Speech Detection.

Challenge competitions	Years	Data types	Languages (number)	Public labels (train & dev/test)	Audio	Visual	Team no.	Top-1 system
ASVspoof 2015 [20]	2015	Speech	English	Yes/Yes	Yes	No	16	Ensemble Model
ASVspoof 2019 (LA Task) [21]	2019	Speech	English	Yes/Yes	Yes	No	48	Ensemble Model
DFDC [15]	2020	Speech	English	Yes/Yes	Yes	Yes	**2114**	Ensemble Model
FTC [22]	2020	Speech	English	No/No	Yes	No	n/a	n/a
ASVspoof 2021 (LA Task) [23]	2021	Speech	English	Yes/Yes	Yes	No	41	Ensemble Model
ASVspoof 2021 (DF Task) [23]	2021	Speech	English	Yes/Yes	Yes	No	33	Ensemble Model
ADD 2022 Track 1 [17]	2022	Speech	Chinese	Yes/Yes	Yes	No	48	Single Model
ADD 2022 Track 2 [17]	2022	Speech	Chinese	Yes/Yes	Yes	No	27	Single Model
ADD 2022 Track 3.2 [17]	2022	Speech	Chinese	Yes/Yes	Yes	No	33	Single Model
ADD 2023 Track 1.2 [24]	2023	Speech	Chinese	No/No	Yes	No	49	Ensemble Model
ADD 2023 Track 2 [24]	2023	Speech	Chinese	No/No	Yes	No	16	Single Model
AV-Deepfake1M [25,26]	2024	Speech	English	Yes/No	Yes	Yes	n/a	n/a
ASVspoof 2024 [27]	2024	Speech	English	Yes/No	Yes	No	53	Ensemble Model
SVDD 2024 [19,28]	2024	Singing	**Multilanguages (6)**	Yes/No	Yes	No	47	Ensemble Model

challenge competitions, which explore fake speech in a conversation, can be separated into two groups. The first group is proposed for only audio [17,20–24,27,29]. Meanwhile, the second group is for video in which a fake video is identified by fake audio, fake image, or both fake audio and image [15,26]. This indicates that DSD is not only treated as an individual task independently but also considered as a sub-task in multimodal systems.

It is also evident that the second group, which focuses on fake video detection, has attracted significantly more research teams (e.g., 2114 teams in the DFDC challenge [15]) compared to the first group (e.g., the largest team count was 74 in the ASVspoof 2021 challenge [23]). This provides an insight that fake video detection is a more compelling task, drawing greater interest and participation from research teams. Regarding top-1 systems in these challenge competitions, they leveraged the ensemble techniques which combine a wide range of input features or multiple models (i.e., most submitted systems mainly use deep learning based models).

Discussion: Given the recent analysis of challenge competitions proposed for the DSD task, some concerns can be indicated. Firstly, the DSD task has drawn attention from the research community and is now recognized as one of the critical components in a complex system of deepfake detection. However, most current challenge competitions are limited to single languages, such as Chinese or English, and primarily focus on detecting fake speech within conversations. Secondly, some challenge competitions have not published datasets for different reasons. For example, FTC [22] was organized by the US government, and the top-performing systems are used by the US government. Similarly, ADD 2023 [24] only provides the dataset for the teams that attended during the competition. These limitations hinder research motivation and further development once the challenges conclude. Third, it is recognized that fake speech utterances are mainly generated from deep-learning-based speech generation systems. Therefore, if selected deep-learning-based speech generators are not general or up-to-date, this significantly affects the effectiveness and quality of the challenge competition. This highlights the need for collaboration between two tasks of deep-learning-based speech generation and detection within the same challenge competition. Competitions like ASVspoof 2024 [27] and ADD 2022 [17] have addressed this by not only publishing datasets but also presenting a two-phase or two-track challenge in which the first phase/track is for Deepfake Speech Generation and the second one is for Deepfake Speech Detection. Finally, regarding techniques used in these competitions, ensemble models have become widely leveraged to enhance performance in many challenge competitions, enabling research teams to develop top-performing systems. However, this approach has several drawbacks, including limited interpretability, increased system complexity, high training costs, and concerns related to power consumption and green AI. Therefore, different aspects of using deep-learning-based models such as using a single model, low complexity, or real-time inference can be regarded as main constraints in challenge competitions for the DSD task in the future. For example, the DCASE challenge Task 1 [30] for Sound Scene Classification requires the submitted systems to obey two constraints: (1) not larger than 128 K parameters and (2) not larger than 30 MMAC units.

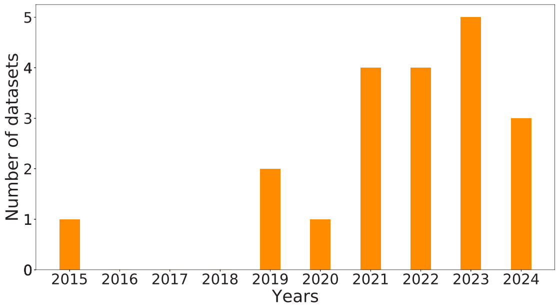

Fig. 3. The number of public datasets proposed for DSD task from 2015.

Conclusion: We have just presented and highlighted the important role of DSD challenge competitions which significantly motivate the DSD research community. We also provided a comprehensive analysis and indicated some existing concerns: (1) the limited number of DSD challenges, particularly for multiple languages; (2) The lack of collaboration between deepfake speech generation and deepfake speech detection; and (3) The absence of real-time or low-complexity requirements for DSD systems in existing challenges. To continue updating new challenge competitions in the future and evaluate the existing concerns, we created a Github project.[3] The GitHub repository serves as a reference for up-to-date information on DSD-related challenge competitions and current concerns. In other words, it complements our survey by ensuring ongoing updates related to the DSD task.

3. Public datasets proposed for deepfake speech detection

Analysis: Public datasets proposed for the DSD task, including those introduced through challenge competitions, play a crucial role in motivating the research community to develop and evaluate DSD systems. In this section, we present a summary of the public and benchmark datasets for the DSD task, as shown in Table 3. These datasets have been introduced through various challenge competitions and published papers.

As illustrated in Fig. 3, the number of public datasets for the DSD task has grown significantly in recent years. Most of these datasets include both clean and noisy speech. Notably, nearly all datasets have

[3] https://github.com/AI-ResearchGroup/A-Comprehensive-Survey-with-Critical-Analysis-for-Deepfake-Speech-Detection

Table 3
Public and benchmark datasets proposed for deepfake speech detection.

Datasets	Years	Languages	Speakers (Male/Female)	Utt. no. (Real/Fake)	Fake speech generators	Speech condition	Real speech resources	Utt. length (s)	Evaluation metrics
ASVspoof 2015 [20] (audio)	2015	English	45/61	16,651/246,500	10	Clean	Speaker Volunteers	1 to 2	EER
FoR [31] (audio)	2019	English	140	-/195541	7	Clean	Kaggle [32]	2.35	Acc
ASVspoof 2019 (LA task) [21] (audio)	2019	English	46/61	12,483/108,978	19	Clean	Speaker Volunteers	n/a	EER
DFDC [15] (video)	2020	English	3426	128,154/104,500	1	Clean & Noisy	Speaker Volunteers	68.8	Pre., Rec.
ASVspoof 2021 (LA task) [23] (audio)	2021	English	21/27	18,452/163,114	13	Clean & Noisy	Speaker Volunteers	n/a	EER
ASVspoof 2021 (DF task) [23] (audio)	2021	English	21/27	**22,617/589,212**	**100+**	Clean & Noisy	Speaker Volunteers	n/a	EER
WaveFake [16] (audio)	2021	English, Japanese	0/2	-/117,985	6	Clean	LJSPEECH [33], JSUT [34]	6/4.8	EER
KoDF [35] (video)	2021	Korean	198/205	62,116/175,776	2	Clean	Speaker Volunteers	90/15 (real/fake)	Acc, AuC
ADD 2022 [17]	2022	Chinese	40/40	3012/24072	2	Clean	AISHELL-3 [36]	1 to 10	EER
FakeAVCeleb [37] (video)	2022	English	250/250	570/25,000	2	Clean & Noisy	Vox-Celeb2 [38]	7	AuC
In-the-Wild [39] (video)	2022	English	58	19963/11816	0	Clean & Noisy	Self-collected	4.3	EER
LAV-DF [40] (video)	2022	English	153	36,431/99,873	1	Clean & Noisy	Vox-Celeb2 [38]	3 to 20	AP
Voc.v [41] (audio)	2023	English	46/61	14,250/41,280	5	Clean & Noisy	ASVspoof 2019	n/a	EER
PartialSpoof [42] (audio)	2023	English	46/61	12,483/108,978	19	Clean & Noisy	ASVspoof 2019	0.2 to 6.4	EER
LibriSeVoc [43] (audio)	2023	English	n/a	13,201/79,206	6	Clean & Noisy	Librispeech	5 to 34	EER
AV-Deepfake1M [25,26] (video)	2023	English	2,068	**286,721/860,039**	2	Clean & Noisy	Voxceleb2 [38]	5 to 35	Acc, AuC
CFAD [44] (audio)	2024	Chinese	1023	-/374,000	11	Clean & Noisy & Codecs	AISHELL1-3 [45,46] MAGICDATA [47]	n/a	EER
MLAAD [48] (audio)	2024	**Multilanguages (23)**	n/a	-/76,000	54	Clean & Noisy	M-AILABS [18]	n/a	Acc
ASVspoof 2024 [27] (audio)	2024	English	n/a	**188,819/815,262**	28	Clean & Noisy	MLS [49]	n/a	EER
SVDD2024 [19] (audio)	2024	**Mutilanguages (6)**	59	12,169/72,235	48	Clean	Mandarin, Japanese	n/a	EER

Table 4
Deepfake speech generation systems used in public DSD datasets (TTS: Text to Speech, VC: Voice Conversion, AT: Adversarial attach using Malafide or Malocopula).

Datasets	Year	No. of TTS/VC/AT	Deepfake speech generation systems
ASVspoof 2015 [20]	2015	7 VC, 3 TTS	VC-01 [50,51], VC-02 [52], TTS-01 [53], TTS-02 [53], VC-03 [54], VC-04 [55], VC-05 [55], VC-06 [56], VC-07 [57], TTS-03 [58]
FoR [31]	2019	7 TTS	Deep Voice 3, Amazon AWS Polly, Baidu TTS, Google Traditional TTS, Google Cloud TTS, Google Wavenet TTS, Microsoft Azure TTS
ASVspoof 2019 (LA task) [21]	2019	8 VC, 11 TTS	TTS-01 [59], TTS-02 [59,60], TTS-03 [61], TTS-04 [62], VC-01 [63], VC-02 [64], TTS-05 [61,65], TTS-06 [59,66], TTS-07 [67,68], TTS-08 [69,70], TTS-09 [69–71], TTS-10 [72], VC-03+TTS [73], VC-04+TTS [74,75], VC-05+TTS [74,75], TTS-11 [62], VC-06 [76,77], VC-07 [78–80], VC-08 [64]
DFDC [15]	2020	1 TTS	TTS Skins voice conversion [81]
KoDF [35]	2021	2 TTS	ATFHP [82] and Wav2Lip [83]
ASVspoof 2021 (LA task) [23]	2021	13 TTS/VC	Reuse ASVspoof 2019
ASVspoof 2021 (DF task) [23]	2021	100 TTS/VC	Vocoders [84]
WaveFake [16]	2021	6 TTS	MelGAN [85], FB-MelGAN [85], HiFi-GAN [86], WaveGlow [87], PWG [88], MB-MelGAN [85]
FakeAVCeleb [37]	2022	2 TTS	SV2TTS [89,90]
In-the-Wild [39]	2022	n/a	n/a
LAV-DF [40]	2022	1 TTS	SV2TTS [89]
Voc.v [41]	2023	5 TTS	HiFi-GAN [86], MB-MelGAN [85], WaveGlow [87], PWG [88], Hn-NSF [91]
PartialSpoof [42]	2023	21 TTS/VC	Reuse ASVspoof 2019
LibriSeVoc [43]	2023	6 TTS/VC	WaveNet [72], WaveRNN [92], MelGAN [85], Parallel WaveGAn [93], WaveGrad [94], DiffWave [95]
AV-Deepfake1M [25,26]	2023	2 TTS	VITS [96], YoursTTS [97]
CFAD [44]	2024	11 TTS	STRAIGHT [98], Griffin-Lim [99], LPCNet [100], WaveNet [72], PWG [88], HiFi-GAN [101], MB-MelGAN [85], MelGAN [85], WORLD [102], FastSpeech [103], Tacotron-HifiGAN [104]
MLAAD [48]	2024	54 TTS	Bark, Capacitron, FastPitch, GlowTTS, Griffin Lim, Jenny, NeuralHMM, Overflow, Parler TTS, Speech5, Tacotron DDC, Tacotron2, Tacotron2 DCA, Tacotron2 DH, Tcotron2-DDC, Tortoise, VITS, VITS Neon, VITS-MMS, XTTS v1.1, XTTS v2
ASVspoof 2024 [27]	2024	15 TTS, 6 VC, 7 AT	TTS-01 [105], TTS-02 [106], TTS-03 [107], TTS-04 [108], TTS-05 [109], TTS-06[110], TTS-07[111], TTS-08(self-develop), VC-01[112], TTS-09[113], VC-02 [114], VC-03(self-develop), TTS-10 [115], AT-01 (Malafide+TTS-10 [115]), TTS-11 [116], AT-02(self-Develop), TTS-12 [117], TTS-13 [118], AT-03(Malafide+TTS [119]), VC-04(self-develop), VC-05 [120], VC-06(add noise), AT-04(Malacopula+VC-06), TTS-14 [121], TTS-15 [122], AT-05(Malacopula+AT-01), AT-06(Malacopula+TTS-13 [118]), AT-07(Malacopula+VC-05 [120])

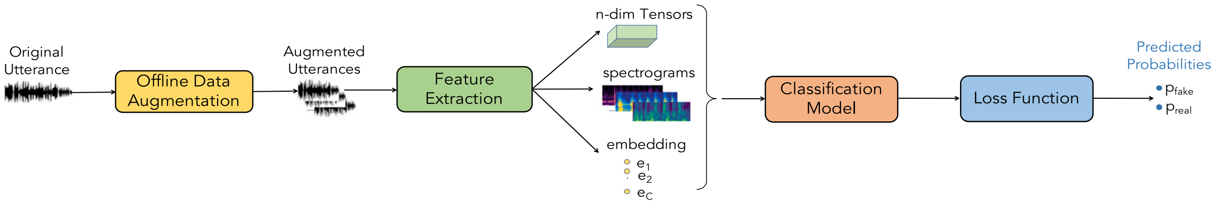

Fig. 4. The high-level architecture of Deepfake Speech Detection (DSD) systems.

been designed for English, with WaveFake [16], KoDF [82], and ADD 2022 [17] being the exceptions, focusing on Japanese, Korean, and Chinese languages, respectively. Recently, the first multilingual datasets for the DSD task were introduced in [18,19]. The MLAAD dataset [18] provides fake speech in conversations generated in 23 widely spoken languages. Meanwhile, the SVDD dataset [19] was proposed for deepfake singing detection with six different languages (i.e., the Chinese songs are the majority).

Most deepfake datasets are generated from one of three generator techniques: Text-to-Speech (TTS), Voice Conversion (VC), and Adversarial Attacks (AT), as shown in Table 4. Notably, ASVspoof 2024 [27] is the first dataset that uses AT systems to generate fake speech. While TTS systems generate fake speech from text, VC systems generate fake speech from real speech (e.g., audio). To mimic the target speakers, TTS and VC systems attempt to explore the audio embeddings extracted from the target speakers. These audio embeddings are treated as a part of the feature map in the entire network architecture in TTS and VC systems. Regarding AT systems, they mainly apply Malafide [123] and Malocopula [124] methods to generate fake speech. Both Malafide [123] and Malocopula [124] methods involve leveraging filter banks. Malafide [123] applies multiple techniques of linear time-invariant (LTI), non-causal filter, and the coefficients (e.g., tap weights) to create TTS/VC-based fake speech that mimics the target speaker. Meanwhile, Malocopula [124] combines both linear filter and non-linear filter (e.g., one-dimensional convolutional layer) to replicate the target speaker's voice.

To compare among DSD datasets, we analyze three different aspects: (1) the number of fake utterances; (2) the AI-synthesized speech systems used to generate fake speech; and (3) the original/real human speech resource used to generate fake speech utterances. As Table 3 shows, most datasets present lower than 300,000 utterances of fake speech, except ASVspoof 2021 (DF Task) [23], ASVspoof 2024 [27], and AV-Deepfake1M dataset [25,26] with 58,9212, 81,5262, and 86,0039 fake utterances, respectively. Although DFDC [15,81] and AV-Deepfake1M dataset [25,26] present a large number of fake data, this was proposed for video in which audio may not be fake. Additionally, these fake utterances were generated from only a few deep-learning-based speech-generation systems. Indeed, two TTS models of VITS [96], YoursTTS [97] and one TTS model [81] were used to generate fake speech in some datasets such as DFDC [15] and AV-Deepfake1M [25, 26], respectively. On the other hand, the ASVspoof 2021 (DF Eva) dataset [23] contains 589,212 fake utterances, generated using over 100 voice conversion (VC) and text-to-speech (TTS) systems. To catch up with state-of-the-art deepfake speech generators, Table 4 presents the architectures and resources of deepfake speech generators. The table indicates that the ASVspoofing series show up-to-date and diverse deepfake speech generators compared to the others. In terms of the original human speech resources, most DSD datasets are based on recordings from a limited number of speaker volunteers. For example, although the ASVspoof 2021 (DF Eva) dataset [23] used 100 VC and TTS systems to create fake utterances, the real speech resource is from 107 speaker volunteers. Some DSD datasets of AV-Deepfake1M [25,26], CFAD [44] leveraged the large and available human speech datasets to generate fake utterances such as Voxceleb2 [38], AISHELLI-3 [36], MAGICDATA [47]. However, these datasets use a limited number of speech generators (e.g., 2 TTS and 11 TTS for AV-Deepfake1M [25,26] and CFAD [44], respectively).

Regarding metrics evaluation, all datasets proposed for the DSD task come together with a baseline and metrics for the evaluation. Regarding the baseline systems, all baselines leveraged convolutional neural network (CNN) based architectures. These baselines are evaluated mainly by the Equal Error Rate (EER) metric. Some datasets such as KoDF [35], AV-Deepfake1M [25,26], MLAAD [48], FoR [31] used Accuracy (Acc.) and Area Under The Curve (AUC) metrics instead of EER.

Discussion : Given the analysis of benchmark datasets proposed for the DSD task, some existing issues can be outlined. These include the limited number of datasets available for multiple languages and the imbalance of several aspects within existing datasets.

Firstly, more public and benchmark datasets have been proposed for the DSD task. However, there is only one multilingual dataset currently. The lack of multilingual datasets for DSD tasks presents several challenges for current model development and evaluation such as performance degradation on cross-language settings that leads to a limited applicability in real-world applications. This motivates the research community to propose more datasets for multiple languages to enhance model's capability in real-life settings. Secondly, another limitation of currently available datasets is that they focus on a limited number of DSD use cases. In particular, two use cases should be clearly distinguished: (1) detecting deepfakes *without* access to the original voice, and (2) detecting deepfakes *with* access to the original voice. The current datasets are designed for addressing the former but not the latter use case as they lack authentic-cloned speech pairing. Another highly relevant use case that should be addressed in the future is *partially* deepfake speech whereby just a part of the speech is being replaced by a synthetic component. Thirdly, we highlight an imbalance among DSD datasets regarding three aspects: (1) the number of fake utterances; (2) the AI-synthesized speech systems used to generate fake speech; and (3) the original/real human speech resource used to generate fake speech utterances. The imbalance can be clearly described in Fig. 5.

- **The number of utterances:** The quantity of utterances within the datasets is not uniform. Some datasets may contain a large number of samples, while others have significantly fewer. A small number of real or fake utterances within datasets (e.g., Fake AVCeleb [37], ADD [17]) limits the model's exposure to a wide variety of speech patterns and scenarios, affecting the detection robustness and generalization on new, unseen data. Additionally, a controlled ratio between real and fake samples created within datasets (e.g., ASVspoof 2024 [27], ASVsproof 2021 [23]) also ensure diversity of fake techniques and avoid overfitting on the fake data, especially if the fake samples are generated using similar techniques. Therefore, maintaining a moderately controlled ratio between real and fake utterances, along with a diverse range of these utterances, is essential for future dataset development.
- **Deepfake speech generation systems:** The variety of deep-learning-based systems used to generate deepfake speech is another area of concern. As Table 4 shows, some of datasets such as MLAAD [48], ASVspoof 2021(DF task) [23], ASVspoof 2024 [27] present more than 20 systems (e.g., TTS, VC, or AT systems).

Table 5
Individual DSD systems exploring raw audio.

Systems	Years	Datasets	Features	Data augmentation (Distoration/Compression)	Models	Loss functions
[125]	2021	ASVspoof 2021 (LA Task)	Raw Audio	Comp.: MP3, ACC, OGG	RawNet2	Focal loss
[126]	2021	ASVspoof 2021 (LA&DF Tasks)	Raw Audio	Comp.: G.723, G.726, GSM, opus, speex, mp2, ogg, tta, wma, acc, ra	RawNet2	Cross Entropy (CE)
[127]	2021	ASVspoof 2019 (LA Task)	Raw Audio	Dis.: Channel Drop, Frequency masking	SinC+CRNN	MSE Loss
[128]	2021	ASVspoof 2021 (LA Task)	Raw Audio	Comp.: mp3, mp2, m4a, m4r, opus, ogg, mov, PCM μ-law, PCM a-law, speex, ilbc, G.729, GSM, G.722, AMR	RawNet2	OC-Softmax
[129]	2021	ASVspoof 2021 (LA&DF Tasks)	Raw Audio	Dis.: Time-wise, Silence Strimming	RawNet2	Cross Entropy
[130]	2021	ASVspoof 2021 (LA&DF Tasks)	Raw Audio	n/a	Encoder: SinC+Residual Decoder: Graph Attention Network	WCE Loss
[131]	2021	ASVspoof 2021 (LA&DF Tasks)	Raw Audio	Dis.: Mixup, FIR filters	Sinc+CNN	WCE Loss
[132]	2021	ASVspoof 2021 (LA Task)	Raw Audio	Comp.: G.711-alaw,G.722, GSM-FR, and G.729	SinC+RawNet2	AM-softmax
[39]	2022	ASVspoof 2019 (LA Task) In The Wild	Raw Audio	n/a	RawNet2, RawNet-GAT, CRNNSpoof	Cross Entropy
[133]	2022	ASVspoof 2019 (LA Task)	Raw Audio	n/a	Encoder: RawNet2 Decoder: Graph Attention Neural Network	WCE Loss
[134]	2022	ASVspoof 2021 (LA&DF Tasks)	Raw Audio	Dis.: RawBoost [135]	Encoder: Sinc+CNN, Wave2Vec2.0+CNN Decoder: Graph Attention network	WCE loss
[136]	2023	ASVspoof 2019 (LA Task), ASVspoof 2021 (LA&DF Tasks)	Raw Audio	Dis.: Stereo speech	Encoder: SinC+ResNet Decoder: Graph Attention network	AM-softmax
[137]	2023	ASVspoof 2019 (LA Task)	Raw Audio	n/a	Encoder: Wav2vec2.0 [138], HuBERT [139] Decoder: LCNN-LSTM-Graph Attention	Cross Entropy
[140]	2023	ADD 2023	Raw Audio	Dis.: Add noise, mix utterance	Encoder: Wav2Vec2.0 Decoder: ECAPA-TDNN	Cross Entropy
[141]	2022	ASVspoof 2019 (LA Task),	Raw Audio	n/a	Encoder: ECAPA-TDNN, RawNet Decoder: Linear layers	Cross Entropy, Triplet loss, AM-Softmax
[142]	2023	ADD 2023	Raw Audio	Dis.:Add noise, vibration, mixup	Encoder: Wav2Vec2.0 Decoder:CNN-Transformer	A-Softmax, Triplet loss, Adversial loss
[143]	2023	ASVspoof 2019 (LA Task), WaveFake, FakeAVCeleb	Raw Audio	n/a	Encoder: Wav2Vec2.0 [138] Decoder: LCNN-Transformer	Triplet, BCE, Adversarial loss
[144]	2024	ASVspoof 2019 (LA Task), ASVspoof 2021 (LA&DF Tasks), In The Wild [39]	Raw Audio	n/a	SincNet/LEAF+ResNet	Cross Entropy
[144]	2024	ASVspoof 2021 (LA&DF Tasks)	Raw Audio	n/a	Encoder: EnCodec [145], AudioDec [146], AudioMAE [147], HuBERT [139], WavLM [148], Whisper [149] Decoder: ResNet	Cross Entropy
[150]	2024	ASVspoof 2019 (LA Task), ASVspoof 2021 (LA&DF Tasks)	Raw Audio	Dis.: Add noise, overlapping	Encoder: WavLM [148], Decoder: Multi-Fusion Attentive	Cross Entropy
[151]	2024	ASVspoof 2019 (LA Task), ASVspoof 2021 (LA Task), In The Wild	Raw Audio	n/a	Encoder: Wav2vec2.0 [138], BEATS [152], LationCLAP [153], AudioCLIP [154], Decoder: Similarity Score Measurement	n/a

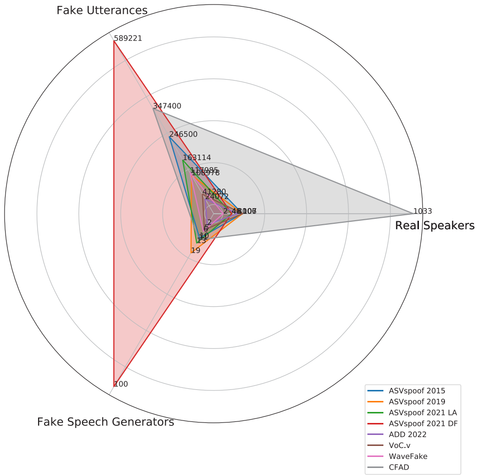

Fig. 5. The imbalance among the fake speech utterances, the fake speech generators, and the real speaker volunteers in benchmark DSD datasets.

Among these datasets, ASVspoof 2021 (DF Task) [23] and ASVspoof 2024 [27] present diverse TTS, VC, and AT systems. In particular, while more than 100 TTS and VC are for ASVspoof 2021 (DF Task) [23], 28 TTS, VC, and AT are used in ASVspoof 2024 [27]. Although MLAAD [48] has been the unique multiple-language dataset currently, fake speech in this dataset was only generated from TTS systems. Overall, some datasets may predominantly feature speech synthesized by a few specific deep-learning-based generators or techniques, while others might include a broader range. Datasets generated from a limited number of deep-learning-based generators possibly lead to over-special ization, reducing the model's ability to detect deepfakes generated by other systems and affecting the performance in real-world scenarios. Therefore, this imbalance motivates the research community to create more diverse datasets that include a wide range of AI-synthesized speech methods.

- **Real human speech resource:** The source of real voice plays a crucial role in shaping the effectiveness, generalization, and ethical aspects of deepfake detection models. As highlighted in Table 4, there are two main sources for building DSD datasets: voice samples from volunteer speakers or from existing datasets. Voice samples from volunteers offer greater control over diversity (if managed thoroughly) and address ethical concerns, as they are collected with explicitly informed consent. However, this approach can be resource-intensive in terms of time and cost and may not scale efficiently. In contrast, utilizing existing human speech datasets offers better accessibility and scalability. However, it may introduce biases toward certain groups, such as public figures, reducing diversity in real-world applications and especially raising significant ethical issues. These problems

Table 6
Individual DSD systems exploring spectrogram based features.

Systems	Years	Datasets	Data augmentation (Distoration/Compression)	Features	Models	Loss functions
[155]	2020	ASVspoof 2019 (LA Task)	Dis.: Add noise, reverberation, FreqAugment	LFCC	ResNet	LMC loss, Cross Entropy
[125]	2021	ASVspoof 2021 (LA Task)	Comp.: MP3, ACC, OGG	LFCC MEL	LCNN TDNN	Focal loss, Focal, Cross Entropy
[156]	2021	ASVspoof 2021 (LA Task)	n/a	LFB, SPEC, LFCC	LCNN, LCNN-LSTM	Cross Entropy, MSE
[157]	2021	ASVspoof 2021 (LA Task)	Comp.: MP3, ACC, landlie, cellular, VoiP	LFCC	ECAPA-TDNN	Focal loss
[126]	2021	ASVspoof 2021 (LA&DF Tasks)	Comp.: G.723, G.726, GSM, opus, speex, mp2, ogg, tta, wma, acc, ra	CQT CQCC, LFCC LFCC	LCNN GMM GMM, LCNN	Cross Entropy
[128]	2021	ASVspoof 2021 (LA Task)	Comp.: G.723, G.726, GSM opus, speex, mp2, ogg, tta, wma, acc, ra	PSCC, LFCC, DCT-DFT, LLFB	Resnet18, TDNN	OC-Softmax
[129]	2021	ASVspoof 2021 (LA&DF Tasks)	Dis.: Time-wise, Silence Strimming	CQT	ResNet, CNN, LSTM	Cross Entropy
[131]	2021	ASVspoof 2021 (LA&DF Tasks)	Dis.: Mixup, FIR filters	MSTFT	ResNet, LCNN	Central loss
[158]	2021	ASVspoof 2019, 2021 (LA Task)	n/a	LFCCs, logLFBs, GM-LFBs, Textrograms	Squeeze CNN	Cross Entropy, A-Softmax loss MLC loss
[159]	2021	ASVspoof 2021 (LA&DF Tasks)	Comp.: MP3, AAC, Landlie, cellular; Dis.: device impulse	LFCCs	ECAPA-TDNN, ResNet	OC-Softmax, P2SGrad losses
[132]	2021	ASVspoof 2021 (LA Task)	Comp.: G.711-alaw, G.722, GSM-FR, and G.729	LFCCs	LCNN	AM-softmax
[160]	2021	ASVspoof 2019 (LA Tasks)	n/a	LFCC	ResNet	OC-Softmax
[161]	2021	ASVspoof 2019 (LA Tasks)	n/a	LFCC	LSTM-SECNN	MSE loss
[162]	2021	ASVspoof 2019 (LA Tasks)	Dis.: SpecAug	log-Mel	ResNet	n/a
[39]	2022	ASVspoof 2019 (LA Task), In the Wild	n/a	CQT, log-STFT MEL	LCNN, CNN-LSTM, Inception, ResNet, Transformer	Cross Entropy
[163]	2022	ADD 2022	Dis.: Add noise/music/babele, Reverb, Modify Volume, SpecAug; Comp.: MP3, OGG, AAC, OPUS	LFCC	ResNet	Focal loss
[164]	2023	ASVspoof 2019 (LA Task), WaveFake, FakeAVCeleb	n/a	LFCC	LCNN-LSTM	Cross Entropy, Adversarial loss, Triplet loss
[165]	2023	ASVspoof 2019 (LA Task)	Comp.: FLAC	MEL	Finetune SSAT Transformer	Cross Entropy
[142]	2023	ASVspoof 2019 (LA Task)	n/a	STFT+F0 sub-bands	SENet34	A-Softmax, KL loss
[166]	2023	ASVspoof 2019 (LA Task)	n/a	LFCC, CQT	Teacher-Student (ResNet, LCNN)	OC-Softmax, MSE loss
[144]	2024	ASVspoof 2019 (LA Task),	n/a	CQT, MEL, logSpec, LFCC	ResNet	Cross Entropy
[167]	2024	ASVspoof 2019 (LA Task), ASVspoof 2021 (LA&DF Tasks)	Dis.: SpecAugment	FBank	ECAPA-TDNN	AM-Softmax
[168]	2024	ASVspoof 2019 (LA Task), ASVspoof 2021 (LA&DF Tasks)	Dis.: RawBoost [135]	log-MEL	Encoder: CNN, ResNet, SE-ResNet Decoder: GAN networks [169]	Cross Entropy, Contrastive loss
[170]	2024	ASVspoof 2019 (LA Task)	Dis.: Oversampling	STFT	Encoder: Transformer Decoder: Transformer	Cross Entropy
[171]	2024	ASVspoof 2019 (LA Task), ASVspoof 2021 (LA&DF Tasks), FakeAVCeleb, WaveFake	Dis.: RawBoost [135]	MEL	Finetune Wav2Vec2.0 (XLSR-53 [138])	Cross Entropy, Contrastive loss
[172]	2024	ASVspoof 2019 (LA Task) ASVspoof 2021 (DF Task)	Comp.: aac, flac, mp3, m4a wma, ogg, wa Dis.: Speed perturbation, SpecAug	LFCC	Encoder: Transformer Decoder: Transformer	OC-Softmax

suggest other balanced approaches to build DSD datasets that consider both diversity and scalability in the future.

Based on the above discussions and statistic information in Fig. 5, it can be concluded that ASVspoof 2019 (LA task) [21], ASVspoof 2021 (LA & DF tasks) [23], ASVspoof 2024 [27] are among the most balanced datasets at the writing time. Additionally, the MLAAD [48] is the largest and most suitable DSD dataset for evaluating cross-languages. The discussions on existing datasets for the DSD task underscore the importance of future efforts by the research community to release comprehensive, multilingual, and balanced datasets. Also, Fig. 5 emphasizes the significant costs and workload involved in creating such datasets, while ensuring compliance with essential security protocols for speaker volunteers.

Conclusion: We have just presented the important role of public datasets proposed for the DSD task, providing a comprehensive analysis and indicating the existing issues. The study shows different aspects that are not mentioned in the other surveys: (1) the original resource of real human speech; (2) the overview of deep learning-based systems used to generate fake speech; (3) not only fake speech but also fake video datasets were mentioned; (4) the imbalances and other concerns in current public DSD datasets, along with their impact on model performance and practical applicability. Similar to the challenges for the DSD task, we will continue to update new DSD datasets via our GitHub repository[4] in the future. This ensures the ongoing relevance of the survey and provides an up-to-date resource for DSD datasets.

4. Overview on proposed systems for deepfake speech detection

To conduct a comprehensive analysis of DSD systems, we first review state-of-the-art research papers addressing the DSD task. Notably, a large number of the selected papers are from high-reputation journals and conferences such as INTERSPEECH (48 papers) and ICASSP (29 papers) in recent years. Then, we categorize these DSD systems into

[4] https://github.com/AI-ResearchGroup/A-Comprehensive-Survey-with-Critical-Analysis-for-Deepfake-Speech-Detection

Table 7
DSD systems leveraging ensemble techniques to enhance the performance.

Systems	Years	Datasets	Features	Data augmentation (Distoration/Compression)	Models	Loss functions	Ensemble methods
[173]	2019	ASVspoof 2019 (LA Task),	LFCC, CQT, FFT	n/a	LCNN	A-Softmax	Multiple inputs
[174]	2021	ASVspoof 2019 (LA Task)	Raw Audio	Dis.: Mixup	ResNet	Cross Entropy	Multiple branches
[175]	2021	ASVspoof 2019 (LA Task)	LSB, SPEC, LFCC	n/a	LCNN, LCNN-LSTM	Cross Entropy, MSE for P2SGrad	Multiple inputs, models
[157]	2021	ASVspoof 2021 (LA&DF Tasks)	LFCC	Comp.: MP3, ACC, landlie, cellular, VoiP	Variants of ECAPA-TDNN	OC-Softmax	Multiple models
[176]	2021	ASVspoof 2021 (LA&DF Tasks)	LFCC	Dis.: Reverberation, add noise, Comp.: mp3, mp4	ResNet, MLP, SWA	large margin cosine, Cross Entropy	Multiple models
[125]	2021	ASVspoof 2021 (LA Task)	LFCC, MFCC, draw	Comp.: MP3, ACC, OGG	TDNN, RawNet2	Focal loss	Multiple inputs, models
[126]	2021	ASVspoof 2021 (LA&DF Tasks)	Draw, CQCC, LFCC	Comp.: G.723, G.726, GSM, opus, speex, mp2, ogg, tta, wma, acc, ra	GMM, LCNN	Cross Entropy	Multiple inputs, models
[128]	2021	ASVspoof 2021 (LA Task)	Raw, PSCC, LFCC, DCT-DFT, LLFB	Comp.: TODO set 1+2	ResNet18, GMM, TDNN, RawNet2	OC-Softmax	Multiple inputs, models
[131]	2021	ASVspoof 2021 (LA Task)	MSTFT	Dis.: Mixup, FIR filters	Resnet18, LCNN, Sinc+CNN	Central loss	Multiple inputs, models
[160]	2021	ASVspoof 2019 (LA Tasks)	LFCC	n/a	ResNet	OC-Softmax	Multiple branches
[177]	2022	ASVspoof 2021 (LA&DF Tasks)	LFCC	Comp.: G.711-alaw, G.711-μlaw	GMM-MobileNet	Cross Entropy	Multiple branches
[178]	2022	ASVspoof 2021 (LA Task)	CQT, MEL	Dis.: Mixup, Frequency Masking	BC-ResNet, FreqCNN	n/a	Multiple inputs, models
[179]	2022	ASVspoof 2019 (LA Tasks)	LFCC	n/a	ResNet, LSTM	OC-Softmax loss	Multiple branches
[180]	2022	ASVspoof 2019, 2021 (LA Task)	Log-Mel	Dis.: Add music, noise, speech Reverb, pitch shift, SpecAug	ResNet	A-Softmax	Multiple models
[181]	2023	ASVspoof 2019, 2021 (LA Task)	Raw Audio	Dis.: Mixup, SpecAug	ResNet	Cross Entropy	Multiple branches
[182]	2023	ADD 2023	Raw Audio, Log-Mel	Dis.: Add noise, room inpulse, mixup, speed shifting, frequency masking	ResNet	Cross Entropy, KL loss	Multiple branches
[137]	2023	ASVspoof 2019 (LA Task)	Wav2vec, Duration, Pronunciation	n/a	LCNN-LSTM-GAP	Cross Entropy Cross Entropy	Multiple inputs
[170]	2024	ASVspoof 2019 (LA Task)	STFT phase, magnitude	Dis.: Oversampling	Transformer	Entropy	Multiple inputs
[183]	2024	ASVspoof 2019 (LA Task), In The Wild	LFCC, MPE	n/a	LCNN	Cross Entropy	Multiple inputs
[184]	2024	ASVspoof 2019 (LA Tasks) ASVspoof 2021 (LA Task) In-the-wild, MLAAD-EN	Raw Audio	Dis.: Noise, Reverb, SpecAug, Drop Frequencies	Encoders: Wav2vec-XLSR-ASR, Wav2vec-XLSR-SER	Cross Entropy, MSE for P2SGrad	Multiple models
[144]	2024	ASVspoof 2019 (LA Task), ASVspoof 2021 (LA&DF Tasks)	Raw Audio	n/a	Encoders: XLS-R, Hubert, WavLM Decoder: ResNet	Cross Entropy	Multiple inputs, models

three groups based on input type, as detailed in Tables 5, 6, and 7. The first group, shown in Table 5, consists of DSD systems that directly process audio utterances using single models. These models are based on a single machine learning algorithm or one specific network architecture. In the second group (Table 6), audio utterances are first transformed into spectrograms, representing temporal-frequency features. After this transformation, a single model is applied to analyze the data. The final group, shown in Table 7, features a diverse range of ensemble models that utilize various input features and combine multiple models.

Given the summary of DSD systems in Tables 5, 6, 7, we describe the high-level architecture of DSD systems as shown in Fig. 4. From Fig. 4, we then identify and analyze four main components that directly impact the DSD system performance: (1) Offline data augmentation, (2) Feature extraction, (3) Classification model, and (4) Loss function and Training strategy.

4.1. Offline data augmentation

Analysis: Data augmentation involves generating variations of the original data to increase the size of DSD datasets, which enhances the robustness and generalization capabilities of machine learning models. Since this step is applied to original audio utterances before the training process, it can be referred to as offline data augmentation. As shown in Tables 5, 6, and 7, offline data augmentation methods can be separated into two main groups, referred to as compression and distortion. The compression methods involve compress and decompress algorithms, mainly using audio codec techniques. A codec, short for 'coder–decoder', is a software used to compress and decompress digital audio. Among these methods, MP3, AAC, OGG, G.7XX, and Opus formats are commonly applied. Codec data augmentation helps simulate these real-world conditions through various compression schemes (e.g., phone calls, music streaming, or online video playback on applications such as Facebook, WhatsApp, etc.). Since different codecs use various compression and decompression algorithms, they impact audio-related factors such as signal-to-noise ratio (SNR), high-frequency formants, energy loss, sample rate, bit depth, and bitrate in distinct ways. This suggests that if there are subtle differences between real and fake speech in these aspects, generating diverse audio utterances using different codecs can be an effective approach for distinguishing between them.

Codec methods can be divided into three main categories based on the quality of audio data: uncompressed format, lossless compressed format, and lossy compressed format. Audio files with uncompressed formats such as WAV, AIFF, or PCM are large and contain all audio information recorded from an audio device. The lossless compressed formats such as FLAC, WMA, or ALAC only reduce unnecessary features of audio data and retain the almost original audio data. Meanwhile, lossy compressed formats such as MP3 or AAC significantly reduce audio features such as sample rate or bit depth to achieve low-volume audio files, which is suitable for streaming-based applications with real-time requirements.

The second distortion method modifies the raw audio by adding reverberation, background noise, and music [176,180,184] or by applying time-wise processing and silence streaming techniques [129], while preserving audio quality parameters such as sample rate, bit depth, and bit rate. The distortion method enforces classification models to learn distinct features between fake and real speech while these features are mixed by different noise resources. Notably, conventional data augmentation methods, such as pitch shifting and time stretching, which are commonly applied to raw audio in tasks like Acoustic Scene Classification [185], Speech Emotion Detection [186], and Speech Separation [187], have not been applied popularly to the DSD task [180, 182].

Discussion: Although compression methods and distortion methods present different approaches to generate more audio data, none of the papers has compared, analyzed, and indicated if one of the approaches is superior in the DSD task. Indeed, the statistical information in Fig. 6 indicates that the number of state-of-the-art DSD systems using

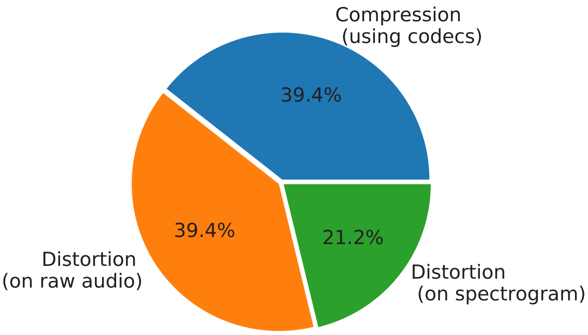

Fig. 6. The statistics of data augmentation methods obtained from Tables 5, 6, 7.

offline distortion augmentation and offline compression augmentation are equal.

Regarding codec-based data augmentation, little research has examined the differences among codec methods to identify which are most suitable for the DSD task in certain real-life scenarios. Indeed, social networks such as Facebook, Instagram, or YouTube and Internet-based communication tools such as WhatsApp, and WeChat (VoIP call) utilize specific and relevant codec methods. For example, YouTube shares audio with MP3 formats, while VoIP calls normally use G.722 audio format as the standard. However, many proposed DSD systems have been evaluated on current and benchmark datasets with WAV files, which do not accurately reflect the codec-specific conditions of real-life DSD applications.

In speech-relevant tasks such as speaker recognition, speaker emotion detection, etc., some distortion data augmentations of Mixup [188] or SpecAugment [189], which are inspired from the computer vision domain, are widely used. These data augmentation methods focus on synthesizing new spectrograms in various manners (e.g., merging, masking), which might not accurately reflect artifacts of the audio signal. Additionally, these data augmentation methods are applied to batches of spectrograms, referred to as online data augmentation. As shown in Fig. 6, Mixup [188] or SpecAugment [189] are also used in a wide range of DSD systems. However, none of the papers has analyzed or compared the efficiency between offline data augmentation and online data augmentation.

Conclusion: Given the analysis and the existing concerns above, we can conclude that although a wide range of data augmentation methods are used, the contribution of each method has not been comprehensively analyzed. Therefore, to evaluate the role and the effect of the online and offline data augmentation methods, we conducted extensive experiments in this paper. Based on our findings, we highlight data augmentation strategies that are most compatible with DSD systems. In particular, we compare the performance of codec-based methods with the Mixup [188] and SpecAugment [189]. On our GitHub repository, we regularly update codec-based methods and other data augmentation techniques featured in the latest research.

4.2. Feature extraction

Analysis: As shown in Fig. 4, feature extraction methods can be categorized into two main groups: non-parameter and trainable-parameter methods. In **non-parameter feature extraction**, a raw audio utterance (e.g., a 1-D tensor) is first transformed into a time-frequency spectral features (e.g., a 2-D tensor) using various transformation ranging from spectral coefficients (e.g., MFCC [125,190], LFCC [128,160,179], CQCC [126], etc.) to spectrogram-based representations such as STFT-spectrogram [131,170], CQT-spectrogram [126,129], etc. Once the time-frequency spectrograms are generated, some DSD systems directly use them for training with classification models [129], while other systems use several approaches to enhance feature quality before applying a classification model. The first approach involves applying auditory filter banks such as Mel [144,157], Linear Filter [125,132,159] (LF), etc, to capture the relationships between frequency bands. Then a Discrete Cosine Transform (DCT) is applied to analyze the relationship across temporal dimension before the features are fed into a model for the training process [125,132,157,159]. Notably, the output of Mel, LF, or DCT operations remains a 2-D tensor (similar to a spectrogram), representing both temporal and spectral features.

In the second approach, audio inputs are fed into pre-trained models, such as XLS-R [191], Hubert [139], WavLM [148], or Whisper [149], to extract embeddings. These embeddings can be the outputs feature maps from specific layers of these pre-trained model [144]. Typically, the embeddings form a 1-D tensor, similar to a vector, where each dimension of the vector is treated as an independent value.

In general, non-parameter feature extraction leverages various spectrogram transformations, auditory filters, auditory statistics, and pre-trained models to generate distinct features (e.g., 1-D audio embeddings, 2-D spectrograms) of audio input.

Trainable-parameter feature extraction involves extracting audio features by applying trainable network layers. In particular, systems proposed in [127,130,144] applied SincNet layers [192], LEAF layers [193], FBanks [167] to learn and extract features from raw audio. These techniques construct learnable filterbanks or approximate the standard filtering process. For example, SincNet and LEAF layers keep the role of adaptive and bandpass filters to capture frequency features between two pre-defined cut-off frequencies. The outputs of these trainable layers are the feature maps that are then fed into the next parts of detection systems. In other words, trainable feature extraction includes trainable network layers as a part of entire network architectures that directly train and learn features from raw audio without the spectrogram transformation steps.

Discussion: By allowing learnable temporal-spatial features during the training process, trainable-parameter feature extraction is compatible with end-to-end systems and shows effectiveness in distinguishing artifacts in fake speech. However, as most proposed systems using trainable features were evaluated on single datasets rather than cross-dataset settings, this possibly leads to challenges in generalization since learned feature sets perform well under specific conditions but fail in unseen fake speech in real-world environments. Regarding feature extraction using audio embeddings from pre-trained models, although these pre-trained models are effective for many audio tasks, using them for deepfake detection presents several challenges. Firstly, as pre-trained models are initially trained for upstream tasks such as speech-to-text, speaker identification, emotion detection, etc, that focus on different aspects (i.e., speech-to-text or emotion detection), the audio melody and harmony (i.e., emotion detection), or distinct frequencies (i.e., speaker identification), embeddings can fail to capture subtle artifacts specific to synthesized speech. Secondly, audio deepfakes are generated to closely mimic real speech, they often have the same formants, pitch, and rhythm as real audio, especially when generated by advanced deep-learning-based speech generation systems. Additionally, the use of pre-trained models can add complexity due to their large network architectures.

For systems using spectrograms such as CQT, MEL, GAM, etc., each spectrogram is designed to capture specific frequency ranges. These spectrograms focus on different central frequencies, which allows them to highlight distinct features of an audio signal. However, human speech contains a wide range of formants - characteristics of sound determined by factors such as language, accent, vocal tract shape, and vocal fold behavior. Therefore, relying on only one type of spectrogram may miss important features, leading to incomplete or insufficient representations of the speech signal that are useful for deepfake detection. To address this, DSD systems have begun to use ensembles of multiple spectrogram inputs [125,126,128,131,157]. By leveraging the unique strengths of each spectrogram type, this approach aims to enhance detection accuracy and has shown significant improvements

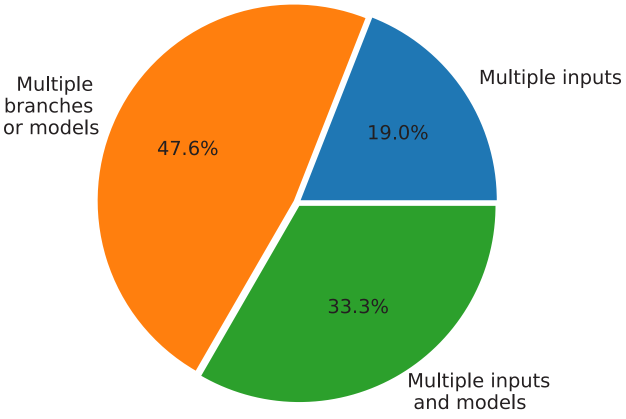

Fig. 7. The statistics of ensemble methods obtained from Tables 5, 6, 7.

in model performance. Many top-performing systems in recent competitions have demonstrated the effectiveness of using ensembles to boost overall system robustness. However, ensemble models present several limitations, including reduced interpretability, increased system complexity, and higher training costs.

Conclusion: We have presented the commonly used feature extraction methods in DSD systems, highlighting their characteristics and potential challenges associated with each approach. In the next section, we conduct extensive experiments of various feature extraction methods to evaluate the most effective approach for the DSD task. Additionally, we explore different feature ensembles to determine the optimal combinations for enhancing performance.

4.3. Classification models

Analysis: Early models proposed for DSD task approached conventional machine learning algorithms. For example, 9 over 16 submitted systems in ASVspoofing 2015 challenge [190] extract MFCC feature (i.e. Systems A, B, E, G, H, I, N, O, and P in [190]). Then, various machine learning-based models such as Mahalanobis distance measurement, Gaussian-based model (GMM), Support vector machine-based models (SVM, SVM-RBF), or fusion models (GMM and SVM) are used to explore MFCC features. Due to the emergence of powerful of deep learning techniques, a wide range of deep neural architectures have been applied to recent DSD systems, as as shown in Tables 5, 6, 7. Recently proposed deep neural networks for the DSD task can be separated into four main approaches. The first approach leverages convolutional-based network architectures (CNNs), which focuses on exploring spatial features. Among the CNN-based networks, Resnet, LCNN, and RawNet architectures are widely used. ResNet and LCNN are used to explore spectrogram-based features such as LFCC [126], CQT [131], and MEL [144]. Meanwhile, RawNet architectures are normally combined with SincNet layer [192] to learn raw audio [125,126,128,129,132,144]. The second approach applies temporal neural network, such as recurrent neural network (RNN) based architectures, which focuses on exploring the temporal features. For example, LSTM-based networks, TDNN, or ECAPA-TDNN are proposed in [125,128,129,157,167], respectively. As shown in Tables 5, 6, 7, RNN-based networks have not been popularly applied for the DSD task compared to the CNN-based architectures. The third approach involves combining both convolutional and temporal network architectures to explore audio features, referred to as hybrid network architectures. In particular, recurrent network-based layers such as LSTM, GRU are combined with CNN-based layers to perform convolutional-recurrent neural network (CRNN) architectures [125,164,167]. The fourth approach utilizes encoder–decoder-based network architectures, which have been widely used for the DSD task and have demonstrated their effectiveness. Apart from the conventional encoder and decoder in transformer-based architectures [170,172], various alternative network architectures have been explored. For instance, encoders based on XLSR-53 [171], WavLM [150], CNN, and ResNet [168] have been investigated as replacements. Regarding decoder architectures, numerous approaches such as GAN-based architectures [168], multi-feature attention [150], and Graph Attention Networks [130,133,134], etc, have been leveraged.

To further enhance the DSD performance, the DSD research community leverages a wide range of ensemble models. These ensemble models can be separated into three main approaches which are marked in the final column in Table 7. In the first approach (Multiple inputs), multiple input features are explored [137,170,173,194]. This approach is inspired by the idea that multiple features contain different and distinct features between fake and real utterances. Given different features, each feature is trained by the same classification model (i.e., the individual model shares the same network architecture but presents different training parameters after the training process). For example, while [170] explores the magnitude and phase features of STFT spectrogram, different features of Wav2Vec embeddings, duration, and pronunciation are explored in [137]. Similarly, multiple spectrograms such as LFCC, CQT, and STFT are trained by one classification model of CNN [195]. Finally, the scores obtained from individual models are fused to achieve the final and best result. The second approach (Multiple branches or models) leverages different network architectures that explore one type of input feature [157,160,174,176,177,180–182,184]. This approach is inspired by the idea that different network architectures are likely to capture distinct properties from the input feature. For example, [176] proposed multiple branches of GMM-DNN and ResNet to explore the LFCC spectrogram. Similarly, [157] explores the raw audio by different variants of ECAPA-TDNN. The final approach (Multiple inputs, models) leverages both multiple input features and different network architectures. For example, [125] explore raw audio by RawNet2. Meanwhile, TDNN and LFCC spectrogram are explored by LCNN. Then, the authors fused three results obtained from three individual models. Similarly, multiple input features of raw audio, CQCC, and LFCC are explored by different models of LCNN, GMM, and RawNet2 in [126]. Ensemble methods are widely adopted in many top-performing systems in DSD challenge competitions.

Discussion: Although many deep neural network architectures have been proposed for the DSD task and evaluated on various benchmark datasets, the best results have been obtained from ensemble methods with multiple inputs or/and different network architectures. The statistics of ensemble models, as shown in Fig. 7, indicate that multiple branches or models are the majority. However, ensemble models present the concern of large trainable parameters. Moreover, none of the research has been analyzed to indicate the individual roles of input features or types of network architectures used in ensemble methods. To demonstrate a robust and general DSD model, the proposed model needs to be evaluated with multiple datasets, cross-datasets, or cross-languages. However, only some recent research [39,143,151,171] evaluated the proposed models with multiple datasets such as ASVspoof 2019 (LA Task), ASVspoof 2021 (LA&DF Task), In The Wild, etc. To the best of our knowledge, none of the research has proposed the evaluation on cross-languages.

Conclusion Given the analysis of feature extraction and classification models above, it can be seen that a wide range of input features have been explored by various classification models. However, none of the papers has made effort to compare different approaches, indicating the potential research directions. Therefore, in this paper, we conduct extensive experiments with various input features, indicating the effective input feature for DSD system performance. We also evaluate a wide range of network architectures leveraging the transfer learning technique, end-to-end training approach, and audio embeddings extracted from state-of-the-art pre-trained models. Given extensive experiments on different input features and various network architectures, we propose an ensemble model that is competitive to state-of-the-art DSD systems.

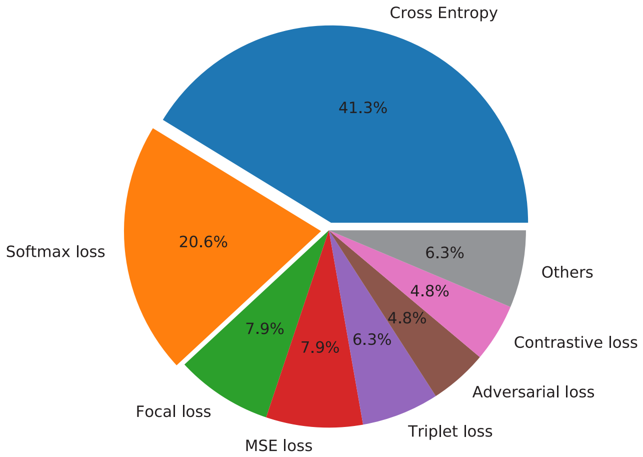

Fig. 8. The statistics of loss functions obtained from Tables 5, 6, 7.

4.4. Loss function and training strategy

From Tables 5, 6, 7, it can be seen that most proposed models use a single loss function. Statistics of the individual loss functions are also presented in Fig. 8. As shown in Fig. 8, the cross entropy (CE) based losses (e.g., Binary Cross Entropy (BCE), Weight Cross Entropy (WCE), etc.) and Softmax-based losses (e.g., Additive-Margin-Based Softmax (AM-Softmax), Angular-Margin-Based Softmax (A-Softmax), etc.) present the most popular loss functions. Some models combine different loss functions. For example, CE and Contrastive loss were used in [168]. Similarly, authors in [164] combined three loss functions of Cross Entropy, Triplet loss, and Adversarial loss. Some papers such as [158,159] compared the DSD performance between large margin cosine loss (LMC loss), and A-Softmax loss functions or between OC-Softmax, MSE for P2SGrad loss functions, respectively.

Generally, a single loss function is used in end-to-end based systems. Meanwhile, the combination of multiple loss functions is related to different training strategies. For example, [171] proposed a teacher-student scheme in which the teacher was trained with contrastive loss and the student was trained by a combination of contrastive loss, Cross Entropy, and MSE loss. Similarly, the student network in [166,196] was trained by a combination of Cosine Similarity/OC-Softmax and MSE loss functions. It can be seen that muliple-loss functions used for teacher-student schemes help achieve a low-complexity model for the DSD task [166,171,196]. Additionally, using multiple-loss function in [141] aims for multiple-task learning strategy. Rather than focusing on loss functions, some researchers improve the DSD system by exploring the training strategy [197–199]. For example, authors in [197] suggested to mix three datasets for the training process. This enhances the generalization and stabilization of the authors' proposed DSD system. Meanwhile, authors in [198] generated more fake utterances by leveraging four types of Vocoders: HiFiGAN, MB-MelGAN, PWG, and WaveGlow, which helps to improve their DSD system performance.

5. Our proposed deepfake speech detection system and extensive evaluation

5.1. Our motivation

Given the comprehensive analysis of the DSD systems in Section 4, we are motivated to conduct extensive experiments that address and evaluate the main concerns below.

- We evaluate the role of offline data augmentation (codec) and compare this method with the conventional online data augmentation methods of Mixup [188] and SpecAugment [189]. We also indicate whether a combination of offline and online data augmentation methods is effective in enhancing the DSD system performance.
- We conduct extensive experiments to evaluate different inputs and network architectures. Given the comparison, we indicate which input features, network architectures, combination of input features, and network architectures have the potential to be further explored. We then propose the best DSD ensemble system that is competitive to the state-of-the-art systems.
- To deeply analyze the role of data augmentation methods, input features, and network architectures, we evaluated proposed DSD systems within cross-dataset and cross-language settings.
- To address the real-time ability, our proposed models are evaluated on two-second utterances and present low-complexity architectures.

5.2. Selected datasets and evaluating metrics

As the trade-off among the number of utterances, the deep-learning-based fake speech generation systems, the original/real human speech resource as shown in Fig. 5 and the comprehensive analysis in Section 3, we decide to use ASVspoof 2019 (LA Task) to evaluate the effect of data augmentations, different types of input features, and various network architectures. Given the results on ASVspoof 2019 (LA Task), we obtain the best DSD systems which are then evaluated with ASVspoof 2021 (LA & DF Tasks) datasets for cross-dataset evaluation and with MLAAD dataset for cross-language evaluation.

We obey the ASVspoof 2019 (LA Task) and ASVspoof 2021 (LA & DF Tasks) challenges, then use the Equal Error Rate (ERR) as the main metric for evaluating proposed models. We also report the Accuracy, F1 score, and AUC score to compare the performance among evaluating models.

5.3. Proposed systems and experimental settings

Data augmentations: We evaluate the role of two data augmentation methods: offline data augmentation (codecs) and online data augmentation (Mixup and SpecAugment). Regarding offline data augmentation using codec-based methods, we use six popular codec formats MP3, OPUS, OGG, GSM, G722, and M4 A. While the codec-based methods compress and decompress raw audio before the training process, the online data augmentation methods of Mixup and SpecAugment work on batches of spectrograms during the training process. By evaluating these two groups of data augmentation individually, we indicate if each of them presents a significant contribution and a combination of two data augmentation methods can help enhance DSD task performance.

Multiple input features: Fig. 9 presents seven types of input features: raw audio and six different spectrograms, which are evaluated in this paper. In particular, we use three transformation methods of Short-time Fourier Transform (STFT), Constant-Q Transform (CQT), and Wavelet Transform. Presumably, each type of spectrogram focuses on different perspectives on frequency content and might catch different inconsistencies in the audio signal. We then leverage different auditory-based filters: Mel and Gammatone filters focus on subtle variations relevant to human auditory perception and the linear filter (LF) isolates specific frequency bands.

As we set the window length, the hop length, and the filter number with 1024, 512, and 64, we achieve the same spectrogram shape of 64 × 64. Then, we apply Discrete Cosine Transform (DCT) to spectrograms across the temporal dimension. Finally, the first and the second-order derivatives are applied to these spectrograms, generating a three-dimensional tensor of 64 × 64 × 3 (i.e., the original spectrogram, the first-order derivative, and the second-order derivative are concatenated across the third dimension).

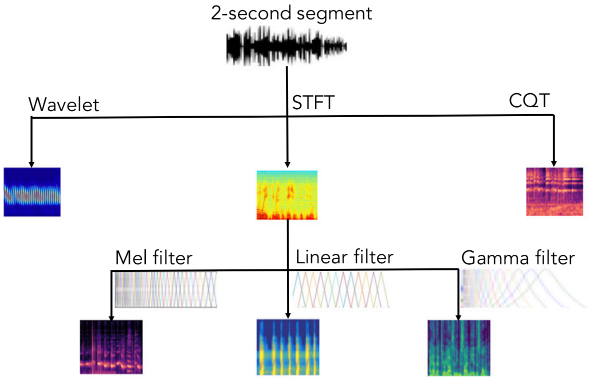

Fig. 9. Generate spectrograms using different spectrogram transformation methods and auditory filter models.

Table 8
The CNN, RNN, and C-RNN network architectures.

Models	Configuration

CNN-based model	**3** × {Conv(32/64/128)-ReLU-AP-Dropout(0.2)}
	1 × {Dense(256)-ReLU-Dropout(0.2)}
	1 × {Dense(2)-Softmax}

RNN-based model	**2** × {BiLSTM(128/64)-ReLU-Dropout(0.2)}
	1 × {Dense(256)-ReLU-Dropout(0.2)}
	1 × {Dense(2)-Softmax}

C-RNN-based model	**3** × {Conv(32/64/128)-ReLU-AP-Dropout(0.2)}
	2 × {BiLSTM(128/64)-ReLU-Dropout(0.2)}
	1 × {Dense(256)-ReLU-Dropout(0.2)}
	1 × {Dense(2)-Softmax}

Table 9
The audio pre-trained models and the Multilayer Perceptron.

Models	Using License	Embedding size/ configuration
Whisper [149]	MIT	512
SpeechBrain [200]	Apache2–0	192
SeamLess [201]	MIT	1024
Pyannote [202,203]	MIT	512
Wav2VecC2.0 [138]	Apache2–0	1024

MLP	Our proposal	**1** × {Dense(128)-ReLU }
		1 × {Dense(2)-Softmax }

Back-end classification models: This paper proposes three main approaches for back-end classification models: the end-to-end deep learning approach, the transfer learning approach, and the audio-embedding deep learning approach. Regarding the end-to-end deep learning approach, four models of CNN-based model, SinC-CNN model (e.g., SinC-CNN architecture is a combination of SinC layer and CNN architecture. The CNN architecture component is reused from CNN-based model), CNN-based model, RNN-based model, and C-RNN-based model are evaluated with the detailed configuration in Table 8. The Sinc-CNN model proves powerful for raw audio input and has been widely used as the survey in Section 4. Meanwhile, CNN-based models are commonly used and effectively capture and learn spectral features. We also use RNNs to focus on detecting natural sequential patterns that can be disrupted in synthetic audio [204] (e.g., temporal coherence, prosodic features such as rhythm, stress, and intonation). Consequently, based on the idea of combining both spectral features and temporal features, we use C-RNN-based model to distinguish characteristics of real and fake audio utterances. With the transfer learning approach, various benchmark network architectures in the computer vision domain are evaluated, such as ResNet-18, MobileNet-V3, EfficientNet-B0, DenseNet-121, SuffleNet-V2, Swint, Convnext-Tiny, GoogLeNet, MNASnet, RegNet, which were trained on the ImageNet1K dataset in advance [205]. Given the pre-trained networks, trainable weights, which capture rich and generalized features of pattern recognition in images, have the potential to adapt patterns in spectrograms by the fine-tuning process. To adapt the DSD task and inspired by [206], we re-use the backbone of the pre-trained models. We then connect the backbone with a dense layer to be compatible with the binary classification task. During the training process, both trainable parameters in the backbone and dense layer are updated with a low learning rate.

For the audio-embedding deep learning approach, different state-of-the-art audio pre-trained models are evaluated. First, we evaluate the Whisper [149] model which was trained for speech-to-text task with multiple languages and supervised training strategy. We also evaluate Speechbrain [200] and Pyannote [202,203] which were proposed for speaker identification task. Finally, we evaluate Seamless [201] and Wav2vec2.0 [138] models which were trained for speech translation and speech-to-text tasks using self-supervised training strategy.

In particular, we feed the spectrogram inputs into these pre-trained models to obtain audio embeddings. Given the audio embeddings, we then propose a Multilayer Perceptron (MLP) to classify these audio embeddings into fake or real classes. The proposed MLP is shown in Table 9, to detect real or fake audio.

Ensemble method: As we train individual model works with two-second audio segments, the result on an entire audio recording is computed by averaging of results over all these segments. Let consider $\boldsymbol{p}^{(n)} = [p_1^{(n)}, p_2^{(n)}, \ldots, p_C^{(n)}]$, where C is the category number of the nth out of N two-second segments, as the predicted probability of one two-second segment. The predicted probability of an entire audio recording, as described by $\bar{\boldsymbol{p}} = [\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_C]$, is computed by:

$$\bar{p}_c = \frac{1}{N}\sum_{n=1}^{N} p_c^{(n)} \quad for \quad 1 \leq c \leq C \tag{1}$$

Given the predicted probabilities from individual models, we propose a MEAN fusion for an ensemble of multiple models. Let consider the predicted probability of one model as $\hat{\mathbf{p}}_s = (\bar{p}_{s_1}, \bar{p}_{s_2}, \ldots, \bar{p}_{s_C})$, where C is the category number and the sth out of S individual models. Next, the predicted probability after MEAN fusion $(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_C)$ is obtained by:

$$\hat{p}_c = \frac{1}{S}\sum_{s=1}^{S} \hat{p}_{s_c} \quad for \quad 1 \leq c \leq C \tag{2}$$

Finally, the predicted label $\hat{y}$ for an entire audio sample is computed by:

$$\hat{y} = \text{argmax}(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_C) \tag{3}$$

5.4. Experimental results and discussions

We first use ASVspoof 2019 (LA Task) to evaluate and indicate the best DSD systems. The comprehensive result comparison is described in Table 10.

Evaluation of data augmentation methods on ASVspoof 2019 (LA Task): Considering the performance of online and offline data augmentation methods as shown in systems A1 (no data augmentation), A2 (online data augmentation with codec), A3 (offline data augmentation with Mixup and SpecAugment), and A4 (both online and offline data augmentation), it can be seen that the offline data augmentations of Mixup and SpecAugment are appropriate for DSD task on ASVspoof 2019 (LA Task) dataset. Notably, the combination of online and offline data augmentations does not help enhance the DSD task performance compared with only using offline data augmentation.

Evaluation of input features on ASVspoof 2019 (LA Task): Considering the efficacy of raw audio and six types of spectrograms in systems from B1 to B7, STFT outperforms the raw audio and other spectrograms. Models B2, B5, and B7 achieve the best ERR score of 0.08

Table 10
Performance comparison among deep learning models on Logic Access evaluation subset in ASVspoofing 2019.

Systems	Inputs	Augmentations	Models	Acc↑	F1↑	AUC↑	ERR ↓
A1	STFT & LF	None	CNN	0.82	0.84	0.91	0.15
A2	STFT & LF	Codec	CNN	0.81	0.84	0.93	0.13
A3	STFT & LF	Mixup, Spec.	CNN	**0.88**	**0.90**	**0.96**	**0.08**
A4	STFT & LF	Codec, Mixup, Spec.	CNN	0.81	0.84	0.93	0.13
B1	Raw Audio	None	SinC-CNN	0.84	0.87	0.96	0.10
B2	STFT	Mixup, Spec.	CNN	**0.87**	**0.89**	**0.96**	**0.08**
B3	CQT	Mixup, Spec.	CNN	0.89	0.90	0.92	0.14
B4	WT	Mixup, Spec.	CNN	0.84	0.86	0.89	0.17
B5	STFT & LF	Mixup, Spec.	CNN	**0.88**	**0.90**	**0.96**	**0.08**
B6	STFT & MEL	Mixup, Spec.	CNN	0.86	0.88	0.95	0.11
B7	STFT & GAM	Mixup, Spec.	CNN	**0.85**	**0.87**	**0.96**	**0.08**
C1	STFT & LF	Mixup, Spec.	RNN	0.92	0.91	0.88	0.17
C2	STFT & LF	Mixup, Spec.	CRNN	0.88	0.90	0.96	0.14
D1	STFT & LF	Mixup, Spec.	ResNet-18	0.49	0.58	0.51	0.47
D2	STFT & LF	Mixup, Spec.	MobileNet-V3	0.59	0.67	0.52	0.48
D3	STFT & LF	Mixup, Spec.	EfficientNet-B0	0.52	0.61	0.51	0.48
D4	STFT & LF	Mixup, Spec.	DenseNet-121	0.58	0.66	0.51	0.48
D5	STFT & LF	Mixup, Spec.	ShuffleNet-V2	0.64	0.71	0.53	0.48
D6	STFT & LF	Mixup, Spec.	Swin_T	**0.84**	**0.87**	**0.94**	**0.09**
D7	STFT & LF	Mixup, Spec.	ConvNeXt-Tiny	**0.88**	**0.90**	**0.96**	**0.075**
D8	STFT & LF	Mixup, Spec.	GoogLeNet	0.53	0.62	0.51	0.47
D9	STFT & LF	Mixup, Spec.	MNASNet	0.62	0.70	0.54	0.47
D10	STFT & LF	Mixup, Spec.	RegNet	0.50	0.60	0.50	0.48
E1	Raw Audio	None	Whisper+MLP	**0.85**	**0.88**	**0.95**	**0.10**
E2	Raw Audio	None	Speechbrain+MLP	0.77	0.81	0.81	0.25
E3	Raw Audio	None	Seamless+MLP	0.86	0.88	0.87	0.20
E4	Raw Audio	None	Pyannote+MLP	0.64	0.71	0.78	0.27
E5	Raw Audio	None	Wav2Vec2.0+MLP	0.79	0.82	0.89	0.18
B2 + B3	STFT, CQT	Mixup, Spec.	CNN	**0.91**	**0.92**	**0.98**	**0.06**
B2 + B4	STFT, WT	Mixup, Spec.	CNN	0.88	0.90	0.96	0.09
B2 + B3 + B4	STFT, CQT, WT	Mixup, Spec.	CNN	0.90	0.92	0.98	0.07
B5 + B6	STFT&LF, STFT&MEL	Mixup, Spec.	CNN	0.88	0.90	0.97	0.08
B5 + B7	STFT&LF, STFT&GAM	Mixup, Spec.	CNN	0.87	0.89	**0.98**	**0.065**
B5 + B6 + B7	STFT& LF, STFT&MEL, STFT&GAM	Mixup, Spec.	CNN	0.88	0.90	0.98	0.069
B5 + D6	STFT&LF	Mixup, Spec.	CNN, Swint_T	0.87	0.89	0.96	0.078
B5 + D7	STFT&LF	Mixup, Spec.	CNN, ConvNeXt-Tiny	0.88	0.90	**0.97**	**0.07**
B5 + D6 + D7	STFT&LF	Mixup, Spec.	CNN, ConvNeXt-Tiny, Swint_T	0.88	0.89	0.97	0.072
B3 + B5 + B7	**CQT, STFT&LF, STFT&GAM**	**Mixup, Spec.**	**CNN**	**0.88**	**0.90**	**0.98**	**0.05**
D7 + E1	**Raw Audio, STFT&LF**	**Mixup, Spec.**	**Whisper, ConvNeXt-Tiny**	**0.86**	**0.88**	**0.99**	**0.03**
B5 + E1	**Raw Audio, STFT&LF**	**Mixup, Spec.**	**Whisper, CNN**	**0.87**	**0.89**	**0.99**	**0.03**

while the combination of STFT & LF obtains slightly better accuracy and F1 scores of 0.88 and 0.9, respectively. This indicates that STFT and applying filters such as Linear Filter or Gammatone filter are suitable for isolating specific frequency bands in classification algorithms.

Evaluation of multiple deep learning approaches on ASV-spoof 2019 (LA Task): Regarding the end-to-end deep learning approach from A1 to C2, CNN systems outperform RNN or C-RNN systems. Indeed, using the same input feature of STFT+LFCC, RNN and C-RNN approaches (C1 and C2 systems) obtain ERR scores of 0.14 and 0.17, which is significantly worse than CNN system (A3 or B2 or B7), with the best score of 0.08. This indicates that the specific patterns indicative of deepfake audio might not be primarily temporal but rather frequency in the spectrogram representation. Regarding the finetuning approach (D1 to D10), Convnext-Tiny stands out as the best system with competitive EER score of 0.075. Meanwhile, the embedding-based approach (E1 to E5) achieves the best EER scores of 0.10 using the pre-trained Whisper model. This suggests the potential of these approaches when choosing the appropriate networks for further optimization.

Evaluation of ensemble methods on ASVspoof 2019 (LA Task): Given the performance of individual input features and network architecture, we conduct extensive experiments to evaluate a wide range of ensemble models. First, ensembles of STFT, CQT, and WT spectrograms are evaluated, indicating the best EER score of 0.06 from the combination of STFT and CQT (B2+B3). Then, ensembles of spectrogram with different filter banks (MEL, LF, GAM) are also evaluated, resulting in the best score of 0.065 from STFT+LF and STFT+GAM (B5+B7). As a result, when an ensemble of CQT, STFT+LF, and STFT+GAM is conducted (B3+B5+B7), we can achieve the EER score of 0.05. Regarding the ensemble of network architectures, CNN and ConvNeXt-Tiny (B5+D7) help obtain the EER score of 0.07. Meanwhile, the combination of Whisper+MLP, ConvNeXt-Tiny (D7+E1) or Whisper+MLP, CNN (B5+E1) achieves the best EER score of 0.03.

We continue evaluating cross-datasets on ASVspoof 2021 (LA & DF Tasks) [23] and cross-languages on MLAAD [18]. For the cross-dataset evaluation, the evaluation sets of ASVspoof 2021 (LA & DF Tasks) [23] are tested with the DSD models which were trained and evaluated on ASVspoof 2019 (LA Task) in advance from Table 10. Regarding cross-language evaluation, we only select pairs of utterances from four languages (e.g., French, Spanish, Italian, and German). A pair of utterances presents the original utterance and a deepfake utterance with the same transcription. Similar to the cross-dataset evaluation, pre-trained DSD systems on ASVspoof 2019 (LA Task) from Table 10 are used to verify the cross-language evaluation.

Table 11
Performance comparison among deep learning models on ASVspoof 2021 (LA &DF Tasks) for cross-dataset evaluation.

Systems	Inputs	Augmentations	Models	Dataset	Acc↑	F1↑	AUC↑	ERR ↓
B5	STFT & LF	Codec	CNN	ASV21-LA	0.84	0.87	0.89	0.16
B5	STFT & LF	Mixup, Spec.	CNN	ASV21-LA	0.88	0.88	0.79	0.27
B5	STFT & LF	Codec & Mixup, Spec.	CNN	ASV21-LA	0.85	0.87	0.90	**0.15**
B5	STFT & LF	Codec	CNN	ASV21-DF	0.88	0.91	0.80	0.27
B5	STFT & LF	Mixup, Spec.	CNN	ASV21-DF	0.91	0.88	0.77	0.28
B5	STFT & LF	Codec & Mixup, Spec.	CNN	ASV21-DF	0.91	0.93	0.80	**0.25**
B3	CQT	Codec	CNN	ASV21-LA	0.76	0.80	0.81	0.23
B3	CQT	Mixup, Spec.	CNN	ASV21-LA	0.73	0.78	0.79	0.26
B3	CQT	Codec & Mixup, Spec.	CNN	ASV21-LA	0.78	0.82	0.82	**0.22**
B3	CQT	Codec	CNN	ASV21-DF	0.71	0.80	0.76	0.29
B3	CQT	Mixup, Spec.	CNN	ASV21-DF	0.68	0.78	0.74	0.31
B3	CQT	Codec & Mixup, Spec.	CNN	ASV21-DF	0.71	0.80	0.77	**0.28**
B7	STFT&GAM	Codec	CNN	ASV21-LA	0.81	0.84	0.86	0.19
B7	STFT&GAM	Mixup, Spec.	CNN	ASV21-LA	0.78	0.82	0.85	0.21
B7	STFT&GAM	Codec & Mixup, Spec.	CNN	ASV21-LA	0.80	0.84	0.85	0.19
B7	STFT&GAM	Codec	CNN	ASV21-DF	0.72	0.81	0.79	0.27
B7	STFT&GAM	Mixup, Spec.	CNN	ASV21-DF	0.73	0.81	0.80	0.27
B7	STFT&GAM	Codec & Mixup, Spec.	CNN	ASV21-DF	0.74	0.82	0.80	**0.26**
D7	STFT & LF	Mixup, Spec.	ConvNeXt-Tiny	ASV21-LA	0.88	0.88	0.73	0.33
E1	Raw Audio	None	Whisper	ASV21-LA	0.84	0.86	0.88	**0.18**
D7	STFT & LF	Mixup, Spec.	ConvNeXt-Tiny	ASV21-DF	0.93	0.94	0.76	0.32
E1	Raw Audio	None	Whisper	ASV21-DF	0.84	0.89	0.92	**0.14**
B5 + E1	Raw Audio, STFT&LF	Mixup, Spec.	Whisper, CNN	ASV21-LA	0.90	0.91	0.96	**0.11**
B5 + E1	Raw Audio, STFT&LF	Mixup, Spec.	Whisper, CNN	ASV21-DF	0.94	0.95	0.95	**0.13**

Evaluation of data augmentation methods for cross-data-set evaluation on ASVspoof 2021 (LA & DF Tasks): As experimental results on the B3, B5, and B7 systems are shown in Table 11, it indicates that using the data augmentation methods helps improve the DSD system performance on both ASVspoof 2021 LA and DF tasks. Significantly, codec shows more effectiveness rather than the online augmentation methods on the ASVspoof 2021 LA task. The results also indicate that a combination of offline data augmenation (e.g., codec) and online data augmentation (e.g., Mixup and SpecAugment) are necessary to achieve a general DSD model to deal with the domain shift issue in cross-data evaluation.

Evaluation of input features for cross-dataset evaluation on ASVspoof 2021 (LA & DF Tasks): Regarding the input features, three types of spectrograms (e.g., CQT, STFT+GAM, STFT+LF) which present the high performance on ASVspoof 2019 dataset are evaluated. In particular, STFT+LF (B5 system) outperforms CQT (B3 system) and STFT+GAM (B7 system). This indicates that a combination of STFT and linear filter is suitable for DSD task.

Evaluation of network architectures for the cross-dataset evaluation on ASVspoof 2021 (LA & DF Tasks): The experimental results from B5 (STFT+LF, CNN), D7 (STFT+LF, ConvNeXt-Tiny) and E1 (Raw Audio, Whisper+MLP) systems indicate that leveraging pre-trained model (E1) significantly outperforms the others. This again proves and explains why more Encoder-Decoder architectures have been recently proposed for the DSD task (i.e., Encoder architectures leveraging pre-trained models such as Whisper or Wave2vec2.0). Regarding the ensemble methods, the combination of B5 and E1, which presents CNN-based architecture and pre-trained Whisper model, achieves the best performance on both ASVspoof 2019 (LA Task) and ASVspoof 2021 (LA & DF Tasks). This also proves that the ensemble of network architectures is more effective than the ensemble of input features.

The results obtained from the evaluation on ASVspoof 2019 (LA Task) and ASVspoof 2021 (LA & DF Tasks) could lead to some conclusions:

- The results indicate a combination of offline data augmentation (codec) and online data augmentation (Mixup, SpecAugment) is essential for constructing a general DSD system.
- Not all network architectures are appropriate for the DSD task. As the good performances obtained from CNN-based network, ConvNeXt-Tiny, Whisper models, it suggests that CNN-based and Encoder-Decoder architectures are suitable for DSD task.
- The ensemble of network architectures is effective in enhancing the model performance on the DSD task rather than the ensemble of spectrograms.
- Leveraging pre-trained models such as Whisper shows effectiveness, reinforcing the growing trend of using Encoder-Decoder architectures with pre-trained Encoders. This explains why these architectures have gained popularity in recent works.

In the cross-language evaluation, as shown in Table 12, all proposed DSD systems exhibit poor performance. This indicates that training a model on a single language (e.g., English) and testing it on other languages (e.g., French, German, Spanish, Italian) is not effective. To develop a robust DSD model for multiple languages, training with multilingual datasets is essential, highlighting the need for the DSD research community to focus on creating and publishing more multilingual datasets for the task.

6. Open challenges and potential research directions

6.1. Datasets for deepfake speech detection

6.1.1. Open challenges

Building better datasets for audio deepfake detection is essential for improving the accuracy and robustness of detection systems. However, the current diversity of available datasets for audio deepfake detection remains limited, especially in terms of speaker identity, language, and deepfake generation methods.

A large number of published datasets feature a narrow range of speaker identities, often focusing on a small group of speakers with

Table 12
Performance comparison among deep learning models on MLAAD dataset for cross-language evaluation.

Systems	Inputs	Augmentations	Models	Dataset-Language	Acc↑	F1↑	AUC↑	ERR ↓
B5	STFT & LF	Codec & Mixup, Spec.	CNN	MLAAD-DE	0.45	0.32	0.53	0.46
B5	STFT & LF	Codec & Mixup, Spec.	CNN	MLAAD-IT	0.49	0.34	0.27	0.69
B5	STFT & LF	Codec & Mixup, Spec.	CNN	MLAAD-FR	0.49	0.35	0.48	0.51
B5	STFT & LF	Codec & Mixup, Spec.	CNN	MLAAD-ES	0.48	0.33	0.45	0.52
E1	Raw Audio	None	Whisper+MLP	MLAAD-DE	0.53	0.52	0.56	0.45
E1	Raw Audio	None	Whisper+MLP	MLAAD-IT	0.52	0.52	0.54	0.48
E1	Raw Audio	None	Whisper+MLP	MLAAD-FR	0.59	0.57	0.62	0.40
E1	Raw Audio	None	Whisper+MLP	MLAAD-ES	0.52	0.52	0.53	0.48
B5 + E1	Raw Audio, STFT & LF	Codec & Mixup, Spec.	CNN, Whisper+MLP	MLAAD-DE	0.50	0.38	0.54	0.47
B5 + E1	Raw Audio, STFT & LF	Codec & Mixup, Spec.	CNN, Whisper+MLP	MLAAD-IT	0.52	0.38	0.63	0.40
B5 + E1	Raw Audio, STFT & LF	Codec & Mixup, Spec.	CNN, Whisper+MLP	MLAAD-FR	0.50	0.36	0.59	0.42
B5 + E1	Raw Audio, STFT & LF	Codec & Mixup, Spec.	CNN, Whisper+MLP	MLAAD-ES	0.50	0.37	0.49	0.50

limited gender, age, and accent diversity. For instance, datasets of ASVspoof and FakeAVCeleb include mainly English-speaking voices from certain groups of speakers (e.g., celebrity, predominantly synthesized voice) with a small number of speakers from different language backgrounds, resulting in biased models when applied to diverse populations.

Many existing datasets are domain-specific, focusing on particular types of audio or speakers. For example, FakeAVCeleb primarily includes celebrity interviews, while LibriSpeech focuses on read recordings. These datasets often have limited variability in terms of recording conditions, speaker interactions, and speech styles, making it difficult to generalize detection models to new domains or unseen environments, such as detecting deepfakes in real-world scenarios with noisy or degraded audio, such as phone calls, public spaces, or online content.

The lack of language diversity is also a significant issue that limits the robustness of detection models. As shown at Table 3, most existing datasets support single languages (primarily English or Chinese). This imbalance raises challenges that hinder the development of robust, audio deepfake detection systems in multilingual settings.

As deepfake generation techniques have been evolving rapidly, they produce fake audio that is increasingly difficult to detect. This makes it difficult for existing datasets to stay up to date as they may be vulnerable to newer methods of audio synthesis. Therefore, datasets must be continuously updated to include samples produced by new techniques to ensure the robustness and adaptability of detection models.

6.1.2. *Future directions*

Given the open challenges discussed in the previous subsection, we highlight some potential future directions in dataset development for Deepfake Speech Detection:

Multilingual and Multimodal Datasets: To address the issue of language diversity, future datasets should include a broader range of languages, accents, and dialects. This variety will enable detection models to better handle diverse linguistic and phonetic features across different languages, ensuring their stability in multilingual contexts and their effectiveness in developing global solutions. Moreover, deepfake content in real-world scenarios often includes both audio and video elements, rather than just audio. Therefore, integrating multimodal datasets that combine both audio and video deepfakes is a crucial direction for future research. This integration enhances detection capabilities by allowing models to identify anomalies across multiple data types, improving their effectiveness in combating increasingly sophisticated forgeries

Continuous Dataset Updates: To stay updated, there needs to be ongoing collaboration between researchers developing deepfake generation methods and those working on the DSD task. Regular updates to datasets should include deepfake samples created by the latest synthesized generation techniques, allowing detection models to adapt to emerging threats.

Cross-Domain and Real-World Dataset Adaptation: One of the biggest challenges for DSD models is domain adaptation — the ability to generalize across different types of audio environments, speakers, and use cases. Future datasets should prioritize cross-domain generalization, including diverse data from various contexts (e.g., podcasts, phone calls, interviews, public speeches, and social media content). In addition, besides varied deepfake generation methods, future dataset development should include data from diverse online platforms (e.g., YouTube, TikTok, podcasts) and various speaker demographics that stimulate inclusive real-life scenarios.

6.2. *The generalization and robustness of deepfake speech detection models*

6.2.1. *Open challenges*

A major challenge in developing deepfake detection systems is ensuring they can generalize to new samples that are not presented in the training data. While models may perform well on known attacks, they often struggle with novel manipulations and across different domains, such as varying languages, accents, or speaking styles. The limited size and diversity of training datasets hinder DSD models' ability to handle real-world variability without degraded performance. Some approaches have been adopted to address these challenges. For example, ensemble models, as discussed in Sections 2 and 4, have been effectively utilized to enhance DSD performance and generalization ability, often achieving top results in competition settings. They are also frequently employed in research papers to deliver competitive outcomes [126,128,144]. While ensemble models are powerful and versatile, they often require significant computational costs during training. Additionally, detection systems leveraging pre-trained models have gained popularity [171]. By fine-tuning models pre-trained on upstream audio tasks like speech-to-text [138,149], the training cost for DSD downstream tasks is greatly reduced. However, proving the generalization of these fine-tuned single models remains challenging. For instance, experiments on ASVspoof 2021 (DF Task) in [171] achieved remarkable results, with an EER of 5.67 compared to 15.64 from the top-performing system in the challenge. In contrast, the performance on the ASVspoof 2021 (LA Task) was much lower, with an EER of 15.92, compared to 1.32 from the top-performing system.

In terms of improving the model's robustness to adversarial attacks, the majority of current methods for defending against adversarial attacks rely on adversarial training [9], which involves generating adversarial examples from known attacks to retrain the model. However, this approach incurs high computational costs.

6.2.2. *Future directions*

To improve the generalization and robustness of detection systems, there has been much room for improving existing approaches as well as proposing new methods. For example, future directions can address challenges in ensemble methods by balancing the trade-off between cost and effectiveness using techniques such as pruning, quantization, and knowledge distillation or other efficient ensembling strategies to reduce model size. In the approach using transfer learning or fine-tuning, employing several strategies such as cross-dataset validation or an ensemble of fine-tuned models could address the challenges of proving generalization. Applying mechanisms to learn information from domain-invariant attacks could also enhance the robustness of models against different adversarial attacks.

6.3. Interpretability and explainable AI (XAI) for deepfake speech detection

6.3.1. Open challenges

Improving interpretability and explainability in Deepfake Speech Detection remains a complex task due to the unique challenges posed by audio data and the black-box nature of deep learning methods. Although various explainable AI (XAI) techniques prove effectiveness in interpreting deep-learning-based models, applying XAI to DSD systems has not drawn much attention from the research community. Indeed, only some recently published papers [207–211] address the role of XAI, which mainly focus on the visualization-based XAI methods. For example, the conventional SHapley Additive exPlanations (SHAP) [212] and Local Interpretable Model-agnostic Explanations (LIME) [213] methods were used to interpret the feature contribution in [208,210] and in [209], respectively. Authors in [207] applied Saliency Map [214] and Smooth Grad [215] techniques to visualize how their model processes audio in the frequency domain. Similarly, layer-wise relevance propagation (LRP), a visualization-based XAI method, was leveraged in [211] to indicate the difference of formants among fake and real audio utterances. While more deep-learning-based models have been proposed to solve the DSD task, not many research papers focus on exploring XAI methods to interpret DSD systems.

6.3.2. Future directions

Based on the above discussion, there is much room for applying XAI to improve transparency and trustworthiness within detection systems. Additionally, leveraging visualization tools for visualizing audio features or feature maps could also provide user-friendly platforms and valuable insights into the underlying decision-making process of detection models.

6.4. Real-time deepfake speech detection

6.4.1. Open challenges

Integrating DSD systems into real-world applications still presents several challenges. Key factors include the length of the audio utterance, the complexity of the model (e.g., the number of trainable parameters), computational costs (e.g., FLOPs), and the target edge devices (e.g., mobile phones, embedded systems, high-performance computers). These factors directly affect inference time and are carefully analyzed to ensure effective implementation. For example, the trade-off between the performance and the model complexity was comprehensively analyzed in [195,216] concerning Acoustic Scene Classification (ASC) task and Acoustic Event Detection (AED) task, respectively. Currently, most proposed DSD systems have been currently evaluated on high-performance computers with the advance of powerful GPUs without any computational constraints, while there is little research on real-time deepfake detection. Several studies, such as [217,218], have proposed real-time deepfake audio detection systems. However, these systems often face significant limitations, such as being applicable to only a limited range of deepfake creation techniques (voice conversion) or domains (communication). These challenges highlight the need for further exploration and analysis of real-time DSD systems in future research.

6.4.2. Future directions

Future directions in developing real-time audio deepfake detection systems could rely on better handling the trade-off between model complexity and performance, facilitating model implementation in low-latency conditions. Some techniques such as quantization and pruning can be used to reduce model size, while other methods leverage edge computing or distributed computing to reduce inference time and handle large-scale data more efficiently.

6.5. Ethical and legal considerations

6.5.1. Open challenges

Training audio deepfake detection models requires large datasets, which may involve the collection and the use of personal voice recordings. For example, VoxCeleb and FakeAV-Celeb corpora contain speech from thousands of celebrities in various environments. Personal data handling raises threats of privacy and consent. Furthermore, there is also a risk of dual-use dilemma when some bad actors could manipulate detection technology and available individuals's speech for harmful purposes such as reinforcing disinformation narratives, defamation, and fraud, infringing on individuals' privacy rights.

6.5.2. Future directions

Future directions in addressing ethical and legal considerations for developing audio deepfake technologies focus on enhancing data privacy protection, fairness, and facilitating global regulatory frameworks. Developers will increasingly incorporate privacy-by-design principles in developing detection systems, ensuring that personal voice data is handled securely and with consent, minimizing the risk of misuse. Within DSD applications, access control mechanisms should be implemented to limit certain groups of people and the frequency of using detection technologies, reducing the potential risk of misuse by malicious actors. In terms of legal perspectives, legal frameworks may also evolve to introduce stricter penalties for misuse of both deepfake creation and detection technology.

6.6. The race between deepfake speech generation and detection

6.6.1. Open challenges

As mentioned and discussed in Section 3, there is a tight relationship between Deepfake Speech Generation and Deepfake Speech Detection tasks. Deepfake Speech Generation systems (e.g., VC, TTS, and AT models) have been becoming more powerful and accessible, enabling the creation of hyper-realistic fake utterances that mimic normal speech patterns and produce fewer detectable flaws. This makes it hard for DSD systems to distinguish between real and manipulated content, presenting challenges to keep pace with these deepfake creation advancements.

6.6.2. Future directions

As deepfakes have evolved rapidly, detection models must also adapt by learning from increasingly realistic fakes. By facilitating collaborative environments, researchers in both Deepfake Speech Generation and Detection can further explore and push boundaries of what is technically possible and ensure that detection methods keep pace with advances in deepfake generators. For example, ADD 2022 [17], ADD 2023 [24], and ASVspoof 2024 [27] challenge competitions were established to engage researchers in both Deepfake Speech Generation and Detection. This promotes innovations in addressing the race between creating and detecting deepfake, improving the robustness of detection systems in combating increasingly complicated deepfakes.

6.7. Feature-free deepfake detection

6.7.1. Open challenges

Deepfake detection faces the usual challenge of the cat-mouse logic of an attack-defense arms race, which is due to the fact that as soon as a feature is identified for detection, it can as quickly be neutralized in the next generation synthesis models. The only way to break this cycle is to develop feature-free detection approaches, which designed systems remain effective against an ever-changing landscape of synthesis techniques, where adaptability and foresight are as critical as accuracy.

6.7.2. Future directions

To develop robust and feature-free deepfake detection approaches, two promising directions can be further explore are leveraging self-supervised models [219,220], which features capturing invariant and high-level features across different types of genuine speech, making them less reliant on artifacts, and using continual learning [221,222], which help remain model's robustness against new emerging attack types. Additionally, another potential method is using the very same synthesis technologies used to produce deepfakes for their own detection. The idea is based on the intuition that an AI model can reproduce speech produced by an AI more easily than by a human, because reality is always more complex than its model. In other words, real speech contains chaotic components that will not be perfectly captured by AI models. The proposed method consists of the training and detection phases. The training phase uses an advanced neural voice cloning system to synthesize voice samples based on the target speech files whose authenticity needs to be verified, and then computes a similarity metric between the target speech (authentic or synthetic) and the cloned speech. This distance distribution is used to find the optimal classification threshold, which is then applied to compute the likelihood of authenticity during the detection phase.

6.8. The availability of deepfake speech detection tools

6.8.1. Open challenges

Deepfake speech detection tools still face challenges in increasing their quantity and quality due to the rapid development of deepfake speech generation techniques. Although DSD systems act as a critical function in Voice over Internet Protocol (VoIP) based platforms such as WhatsApp, Facebook, etc. or social media such as YouTube, Twister, etc. for a thread warning, very few VoIP platforms or social media have announced an available and independent DSD tool. Regarding non-commercial or commercial solutions, only some DSD tools or platforms such as Deepware, WeVerify, TrueMedia, and DeepFake-O-Meter are available as highlighted in the survey [223]. However, information on DSD models used in these tools has been not described in detail except TrueMeida and DeepFake-O-Meter with 3 and 5 systems replicated from published papers. Overall, the sufficiency of deepfake detection applications is primarily due to technical complexity in developing and updating models, resource demands such as computational costs and scalability, accuracy concerns, and privacy issues.

6.8.2. Future directions

To address the mentioned challenges, future improvements in developing deepfake speech detection tools could rely on some approaches such as lightweight detection models that can operate on consumer devices such as smartphones, laptops, or cloud-based services. To ensure broader adaption, the development of open-source deepfake detection tools or libraries and established standards for their use could also be promoted by the collaboration between tech companies and academic institutions, making detection tools more accessible and reliable.

7. Conclusion

This paper has provided a comprehensive survey for Deepfake Speech Detection (DSD) task by deeply analyzing the challenge competitions, the public and benchmark datasets, the main components in a deep-learning-based DSD system. From the survey, we indicate exiting concerns and provide enhance solutions to motivate the research community for further contribution on this research topic. More than a survey, we verified the role and the effect of data augmentation, feature extraction, and network architectures. Given the comprehensive survey and extensive experiments, we indicate potential and promising research directions for Deepfake Speech Detection task.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Lam Pham reports financial support was provided by Austrian Institute of Technology GmbH. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work described in this paper is performed in the H2020 project STARLIGHT ("Sustainable Autonomy and Resilience for LEAs using AI against High priority Threats"). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101021797.

The work in this paper has further received funding from the European Union - European Defence Fund under GA no. 101121418 (EUCINF). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Defame Fakes is funded by the Austrian security research programme KIRAS of the Federal Ministry of Finance (BMF).

Data availability

Data will be made available on request.

References

[1] Z. Khanjani, G. Watson, V.P. Janeja, How deep are the fakes? focusing on audio deepfake: A survey, 2021, arXiv preprint arXiv:2111.14203.

[2] M. Masood, M. Nawaz, K.M. Malik, A. Javed, A. Irtaza, H. Malik, Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward, Appl. Intell. 53 (4) (2023) 3974–4026.

[3] R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan, S. Parkinson, A survey on the detection and impacts of deepfakes in visual, audio, and textual formats, IEEE Access 11 (2023) 144497–144529.

[4] Y. Patel, S. Tanwar, R. Gupta, P. Bhattacharya, I.E. Davidson, R. Nyameko, S. Aluvala, V. Vimal, Deepfake generation and detection: Case study and challenges, IEEE Access 11 (2023) 143296–143323.

[5] J. Yi, C. Wang, J. Tao, X. Zhang, C.Y. Zhang, Y. Zhao, Audio deepfake detection: A survey, 2023, arXiv preprint arXiv:2308.14970.

[6] Z. Khanjani, G. Watson, V.P. Janeja, Audio deepfakes: A survey, Front. Big Data 5 (2023) 1001063.

[7] Z. Akhtar, T.L. Pendyala, V.S. Athmakuri, Video and audio deepfake datasets and open issues in deepfake technology: Being ahead of the curve, Forensic Sci. 4 (3) (2024) 289–377.

[8] E. Altuncu, V.N. Franqueira, S. Li, Deepfake: definitions, performance metrics and standards, datasets, and a meta-review, Front. Big Data 7 (2024) 1400024.

[9] M. Li, Y. Ahmadiadli, X.-P. Zhang, Audio anti-spoofing detection: A survey, 2024, arXiv preprint arXiv:2404.13914.

[10] J. Wu, W. Gan, Z. Chen, S. Wan, H. Lin, Ai-generated content (aigc): A survey, 2023, arXiv preprint arXiv:2304.06632.

[11] B. Yetiştiren, I. Özsoy, M. Ayerdem, E. Tüzün, Evaluating the code quality of ai-assisted code generation tools: An empirical study on github copilot, Amazon codewhisperer, and Chatgpt, 2023, arXiv preprint arXiv:2304.10778.

[12] X. Tan, T. Qin, F. Soong, T.-Y. Liu, A survey on neural speech synthesis, 2021, arXiv preprint arXiv:2106.15561.

[13] B. Sisman, J. Yamagishi, S. King, H. Li, An overview of voice conversion and its challenges: From statistical modeling to deep learning, IEEE/ ACM Trans. Audio Speech Lang. Process. 29 (2021) 132–157.

[14] F. Dakalbab, M.A. Talib, O.A. Waraga, A.B. Nassif, S. Abbas, Q. Nasir, Artificial intelligence & crime prediction: A systematic literature review, Soc. Sci. Humanit. Open 6 (1) (2022) 100342.

[15] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C.C. Ferrer, The deepfake detection challenge (DFDC) dataset, 2020, arXiv preprint arXiv:2006.07397.

[16] J. Frank, L. Schönherr, Wavefake: A data set to facilitate audio deepfake detection, in: NeurIPS, 2024.

[17] Audio deep synthesis detection challenge (ADD 2022), 2022, http://addchallenge.cn/add2022.

[18] M-AILABS speech dataset, 2024, https://github.com/imdatceleste/m-ailabs-dataset.

[19] Y. Zhang, Y. Zang, J. Shi, R. Yamamoto, T. Toda, Z. Duan, SVDD 2024: The inaugural singing voice deepfake detection challenge, 2024, arXiv preprint arXiv:2408.16132.

[20] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, A. Sizov, ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge, in: Proc. INTERSPEECH, 2015, pp. 2037–2041.

[21] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K.A. Lee, et al., ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech, Comput. Speech Lang. 64 (2020) 101114.

[22] The FTC voice cloning challenge, 2023, https://www.ftc.gov/news-events/contests/ftc-voice-cloning-challenge.

[23] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K.A. Lee, T. Kinnunen, N. Evans, et al., ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection, in: Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge, ASVspoof, 2021.

[24] Audio deep synthesis detection challenge (ADD 2023), 2023, http://addchallenge.cn/add2023.

[25] Z. Cai, S. Ghosh, A.P. Adatia, M. Hayat, A. Dhall, K. Stefanov, AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset, 2023, arXiv preprint arXiv:2311.15308.

[26] 1M-deepfakes detection challenge, 2023, https://deepfakes1m.github.io/.

[27] The ASVspoof 2024 challenge, 2024, https://www.asvspoof.org/.

[28] The singing voice deepfake detection challenge (SVDD), 2024, https://challenge.singfake.org/.

[29] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K.A. Lee, J. Yamagishi, ASVspoof 2017 version 2.0: meta-data analysis and baseline enhancements, in: The Speaker and Language Recognition Workshop, 2018, pp. 296–303.

[30] DCASE Challenge Committee, DCASE 2022 challenge - task 1A: Low-complexity acoustic scene classification, 2022, https://dcase.community/challenge2022/task-low-complexity-acoustic-scene-classification.

[31] R. Reimao, V. Tzerpos, FoR: A dataset for synthetic speech detection, in: International Conference on Speech Technology and Human-Computer Dialogue, 2019, pp. 1–10.

[32] Audio source used to generate FoR dataset, 2018, https://www.kaggle.com/datasets/percevalw/englishfrench-translations.

[33] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, K. Kavukcuoglu, Efficient neural audio synthesis, in: Proc. ICML, 2018, pp. 2410–2419.

[34] R. Sonobe, S. Takamichi, H. Saruwatari, JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis, 2017, arXiv preprint arXiv:1711.00354.

[35] P. Kwon, J. You, G. Nam, S. Park, G. Chae, Kodf: A large-scale korean deepfake detection dataset, in: Proc. IEEE/CVF International Conference on Computer Vision, 2021, pp. 10744–10753.

[36] Y. Shi, H. Bu, X. Xu, S. Zhang, M. Li, AISHELL-3: A multi-speaker mandarin TTS corpus, in: Proc. INTERSPEECH, 2021, pp. 2756–2760.

[37] H. Khalid, S. Tariq, M. Kim, S.S. Woo, FakeAVCeleb: A novel audio-video multimodal deepfake dataset, in: Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

[38] J.S. Chung, A. Nagrani, A. Zisserman, VoxCeleb2: Deep speaker recognition, in: Proc. INTERSPEECH, 2018, pp. 1086–1090.

[39] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, K. Böttinger, Does audio deepfake detection generalize? in: Proc. INTERSPEECH, 2022, pp. 2783–2787.

[40] Z. Cai, K. Stefanov, A. Dhall, M. Hayat, Do you really mean that? Content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization, in: International Conference on Digital Image Computing: Techniques and Applications, 2022, pp. 1–10.

[41] X. Wang, J. Yamagishi, Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders, in: Proc. ICASSP, 2023, pp. 1–5.

[42] L. Zhang, X. Wang, E. Cooper, N. Evans, J. Yamagishi, The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance, IEEE/ ACM Trans. Audio Speech Lang. Process. 31 (2022) 813–825.

[43] C. Sun, S. Jia, S. Hou, S. Lyu, Ai-synthesized voice detection using neural vocoder artifacts, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 904–912.

[44] H. Ma, J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang, R. Fu, CFAD: A Chinese dataset for fake audio detection, Speech Commun. 164 (2024) 103122.

[45] H. Bu, J. Du, X. Na, B. Wu, H. Zheng, AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline, in: Proc. O-COCOSDA, 2017, pp. 1–5.

[46] Y. Shi, H. Bu, X. Xu, S. Zhang, M. Li, AISHELL-3: A multi-speaker Mandarin TTS corpus, in: Proc. INTERSPEECH, 2021, pp. 2756–2760.

[47] Z. Yang, Y. Chen, L. Luo, R. Yang, L. Ye, G. Cheng, J. Xu, Y. Jin, Q. Zhang, P. Zhang, L. Xie, Y. Yan, Open source MagicData-RAMC: A rich annotated Mandarin conversational(RAMC) speech dataset, in: Proc. INTERSPEECH, 2022, pp. 1736–1740.

[48] N.M. Müller, P. Kawa, W.H. Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, K. Böttinger, MLAAD: The multi-language audio anti-spoofing dataset, in: Proc. IJCNN, 2024.

[49] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, R. Collobert, MLS: A large-scale multilingual dataset for speech research, in: Proc. INTERSPEECH, 2020, pp. 2757–2761.

[50] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, Y. Stylianou, Towards a voice conversion system based on frame selection, in: Proc. ICASSP, 2007, pp. IV–513.

[51] Z. Wu, T. Virtanen, T. Kinnunen, E. Chng, H. Li, Exemplar-based unit selection for voice conversion utilizing temporal information, in: Proc. INTERSPEECH, 2013, pp. 3057–3061.

[52] T. Fukuda, An adaptive algorithm for mel-cepstral analysis of speech, in: Proc. ICASSP, 1992, pp. 137–140.

[53] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, J. Isogai, Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm, IEEE Trans. Audio Speech Lang. Process. 17 (1) (2009) 66–83.

[54] Festvox voice conversion system, 2024, http://www.festvox.org.

[55] T. Toda, A.W. Black, K. Tokuda, Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory, IEEE Trans. Audio Speech Lang. Process. 15 (8) (2007) 2222–2235.

[56] D. Saito, K. Yamamoto, N. Minematsu, K. Hirose, One-to-many voice conversion based on tensor representation of speaker space, in: Proc. INTERSPEECH, 2011, pp. 653–656.

[57] E. Helander, H. Silén, T. Virtanen, M. Gabbouj, Voice conversion using dynamic kernel partial least squares regression, IEEE Trans. Audio Speech Lang. Process. 20 (3) (2011) 806–817.

[58] MaryTTS speech synthesis system, 2024, http://mary.dfki.de.

[59] HTS working group, the English TTS system Flite+HTS engine, 2014, http://hts-engine.sourceforge.net/.

[60] M. Morise, F. Yokomori, K. Ozawa, World: a vocoder-based high-quality speech synthesis system for real-time applications, IEICE Trans. Inf. Syst. 99 (7) (2016) 1877–1884.

[61] Z. Wu, O. Watts, S. King, Merlin: An open source neural network speech synthesis system, in: Speech Synthesis Workshop, 2016, pp. 202–207.

[62] M. Schröder, M. Charfuelan, S. Pammi, I. Steiner, Open source voice creation toolkit for the MARY TTS platform, in: Proc. INTERSPEECH, 2011, pp. 3253–3256.

[63] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, H.-M. Wang, Voice conversion from non-parallel corpora using variational auto-encoder, in: Proc. APSIPA, 2016, pp. 1–6.

[64] D. Matrouf, J.-F. Bonastre, C. Fredouille, Effect of speech transformation on impostor acceptance, in: Proc. ICASSP, vol. 1, 2006, p. I.

[65] K. Tanaka, H. Kameoka, T. Kaneko, N. Hojo, WaveCycleGAN2: Time-domain neural post-filter for speech waveform generation, 2019, arXiv preprint arXiv:1904.02892.

[66] X. Wang, S. Takaki, J. Yamagishi, Neural source-filter-based waveform model for statistical parametric speech synthesis, in: Proc. ICASSP, 2019, pp. 5916–5920.

[67] H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson, P. Szczepaniak, Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices, in: Proc. INTERSPEECH, 2016, pp. 2273–2277.

[68] Y. Agiomyrgiannakis, Vocaine the vocoder and applications in speech synthesis, in: Proc. ICASSP, 2015, pp. 4230–4234.

[69] L. Wan, Q. Wang, A. Papir, I.L. Moreno, Generalized end-to-end loss for speaker verification, in: Proc. ICASSP, 2018, pp. 4879–4883.

[70] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, K. Kavukcuoglu, Efficient neural audio synthesis, in: Proc. ICML, 2018, pp. 2410–2419.

[71] D. Griffin, J. Lim, Signal estimation from modified short-time Fourier transform, IEEE Trans. Acoust. Speech Signal Process. 32 (2) (1984) 236–243.

[72] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, WaveNet: A generative model for raw audio, in: Proc. Workshop on Speech Synthesis, 2016, p. 125.

[73] Voicetext, 2024, http://dws2.voicetext.jp/tomcat/demonstration/top.html.

[74] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, L.-R. Dai, WaveNet vocoder with limited training data for voice conversion, in: Proc. INTERSPEECH, 2018, pp. 1983–1987.

[75] H. Kawahara, I. Masuda-Katsuse, A. De Cheveigne, Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds, Speech Commun. 27 (3–4) (1999) 187–207.

[76] K. Kobayashi, T. Toda, S. Nakamura, Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential, Speech Commun. 99 (2018) 211–220.

[77] W.-C. Huang, Y.-C. Wu, K. Kobayashi, Y.-H. Peng, H.-T. Hwang, P.L. Tobing, Y. Tsao, H.-M. Wang, T. Toda, Generalization of spectrum differential based direct waveform modification for voice conversion, in: Proc. Workshop on Speech Synthesis, 2019, pp. 57–62.

[78] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Trans. Audio Speech Lang. Process. 19 (4) (2010) 788–798.

[79] P. Kenny, A small footprint i-vector extractor, in: Odyssey, vol. 2012, 2012, pp. 1–6.

[80] S.J. Prince, J.H. Elder, Probabilistic linear discriminant analysis for inferences about identity, in: Proc. IEEE International Conference on Computer Vision, 2007, pp. 1–8.

[81] A. Polyak, L. Wolf, Y. Taigman, TTS Skins: Speaker Conversion via ASR, in: Proc. INTERSPEECH, 2020, pp. 786–790.

[82] R. Yi, Z. Ye, J. Zhang, H. Bao, Y.-J. Liu, Audio-driven talking face video generation with learning-based personalized head pose, 2020, arXiv preprint arXiv:2002.10137.

[83] K. Prajwal, R. Mukhopadhyay, V.P. Namboodiri, C. Jawahar, A lip sync expert is all you need for speech to lip generation in the wild, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 484–492.

[84] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, K.A. Lee, ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild, IEEE/ ACM Trans. Audio Speech Lang. Process. 31 (2023) 2507–2522.

[85] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W.Z. Teoh, J. Sotelo, A. De Brebisson, Y. Bengio, A.C. Courville, Melgan: Generative adversarial networks for conditional waveform synthesis, Adv. Neural Inf. Process. Syst. 32 (2019) (2019).

[86] J. Kong, J. Kim, J. Bae, Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, Adv. Neural Inf. Process. Syst. 33 (2020) 17022–17033.

[87] D.P. Kingma, P. Dhariwal, Glow: Generative flow with invertible 1x1 convolutions, Adv. Neural Inf. Process. Syst. 31 (2018) (2018).

[88] R. Yamamoto, E. Song, J.-M. Kim, Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram, in: Proc. ICASSP, 2020, pp. 6199–6203.

[89] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu, et al., Transfer learning from speaker verification to multispeaker text-to-speech synthesis, Adv. Neural Inf. Process. Syst. 31 (2018) (2018).

[90] K. Prajwal, R. Mukhopadhyay, V.P. Namboodiri, C. Jawahar, A lip sync expert is all you need for speech to lip generation in the wild, in: Proc. ACM International Conference on Multimedia, 2020, pp. 484–492.

[91] X. Wang, S. Takaki, J. Yamagishi, Neural source-filter waveform models for statistical parametric speech synthesis, IEEE/ ACM Trans. Audio Speech Lang. Process. 28 (2019) 402–415.

[92] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, K. Kavukcuoglu, Efficient neural audio synthesis, in: Proc. ICML, 2018, pp. 2410–2419.

[93] R. Yamamoto, E. Song, J.-M. Kim, Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram, in: Proc. ICASSP, 2020, pp. 6199–6203.

[94] N. Chen, Y. Zhang, H. Zen, R.J. Weiss, M. Norouzi, W. Chan, WaveGrad: Estimating gradients for waveform generation, in: Proc. ICLR, 2021.

[95] Z. Kong, W. Ping, J. Huang, K. Zhao, B. Catanzaro, DiffWave: A versatile diffusion model for audio synthesis, in: Proc. ICLR, 2021.

[96] J. Kim, J. Kong, J. Son, Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech, in: Proc. ICML, 2021, pp. 5530–5540.

[97] E. Casanova, J. Weber, C.D. Shulby, A.C. Junior, E. Gölge, M.A. Ponti, Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone, in: Proc. ICML, 2022, pp. 2709–2720.

[98] H. Kawahara, STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds, Acoust. Sci. Technol. 27 (6) (2006) 349–353.

[99] N. Perraudin, P. Balazs, P.L. Søndergaard, A fast griffin-lim algorithm, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2013, pp. 1–4.

[100] J.-M. Valin, J. Skoglund, LPCNET: Improving neural speech synthesis through linear prediction, in: Proc. ICASSP, 2019, pp. 5891–5895.

[101] J. Kong, J. Kim, J. Bae, HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis, in: Proc. NeurIPS, 2020.

[102] M. Morise, F. Yokomori, K. Ozawa, WORLD: A vocoder-based high-quality speech synthesis system for real-time applications, IEICE Trans. Inf. Syst. E99.D (7) (2016) 1877–1884.

[103] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, T.-Y. Liu, Fastspeech 2: Fast and high-quality end-to-end text to speech, 2020, arXiv preprint arXiv:2006.04558.

[104] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R.J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, R.A. Saurous, Tacotron: Towards end-to-end speech synthesis, in: Proc. INTERSPEECH, 2017, pp. 4006–4010.

[105] J. Kim, S. Kim, J. Kong, S. Yoon, Glow-tts: A generative flow for text-to-speech via monotonic alignment search, Adv. Neural Inf. Process. Syst. 33 (2020) 8067–8077.

[106] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, Grad-tts: A diffusion probabilistic model for text-to-speech, in: Proc. ICML, 2021, pp. 8599–8608.

[107] A. Łańcucki, Fastpitch: Parallel text-to-speech with pitch prediction, in: Proc. ICASSP, 2021, pp. 6588–6592.

[108] J. Kim, J. Kong, J. Son, Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech, in: Proc. ICML, 2021, pp. 5530–5540.

[109] F. Lux, J. Koch, N. Thang Vu, Low-resource multilingual and zero-shot multispeaker TTS, in: Proc. AACL, 2022, pp. 741–751.

[110] J. Kong, J. Kim, J. Bae, Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, Adv. Neural Inf. Process. Syst. 33 (2020) 17022–17033.

[111] J. Shen, R. Pang, R.J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al., Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, in: Proc. ICASSP, 2018, pp. 4779–4783.

[112] Y.A. Li, A. Zare, N. Mesgarani, Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion, in: Proc. INTERSPEECH, 2021, pp. 1349–1353.

[113] E. Casanova, J. Weber, C.D. Shulby, A.C. Junior, E. Gölge, M.A. Ponti, Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone, in: Proc. ICML, 2022, pp. 2709–2720.

[114] E.A. AlBadawy, S. Lyu, Voice conversion using speech-to-speech neuro-style transfer, in: Proc. INTERSPEECH, 2020, pp. 4726–4730.

[115] C. Gong, X. Wang, E. Cooper, D. Wells, L. Wang, J. Dang, K. Richmond, J. Yamagishi, Zmm-tts: Zero-shot multilingual and multispeaker speech synthesis conditioned on self-supervised discrete speech representations, IEEE/ ACM Trans. Audio Speech Lang. Process. (2024).

[116] I. Steiner, S.L. Maguer, Creating new language and voice components for the updated MaryTTS text-to-speech synthesis platform, in: Proc. LREC, 2018, pp. 1371–1375.

[117] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, S. Yoon, Bigvgan: A universal neural vocoder with large-scale training, in: Proc. ICLR, 2022.

[118] F. Lux, J. Koch, N.T. Vu, Exact prosody cloning in zero-shot multispeaker text-to-speech, in: Proc. SLT, 2023, pp. 962–969.

[119] F. Lux, J. Koch, N.T. Vu, Low-resource multilingual and zero-shot multispeaker TTS, in: Proc. AACL, 2022.

[120] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, J. Wei, Diffusion-based voice conversion with fast maximum likelihood sampling scheme, in: Proc. ICLR, 2022.

[121] E. Casanova, J. Weber, C.D. Shulby, A.C. Junior, E. Gölge, M.A. Ponti, Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone, in: Proc. ICML, 2022, pp. 2709–2720.

[122] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, et al., XTTS: a massively multilingual zero-shot text-to-speech model, in: Proc. INTERSPEECH, 2024, pp. 4978–4982.

[123] M. Panariello, W. Ge, H. Tak, M. Todisco, N. Evans, Malafide: a novel adversarial convolutive noise attack against deepfake and spoofing detection systems, in: Proc. INTERSPEECH, 2023, pp. 2868–2872.

[124] M. Todisco, M. Panariello, X. Wang, H. Delgado, K.A. Lee, N. Evans, Malacopula: adversarial automatic speaker verification attacks using a neural-based generalised Hammerstein model, 2024, arXiv preprint arXiv:2408.09300.

[125] J. Cáceres, R. Font, T. Grau, J. Molina, B.V. SL, The biometric vox system for the ASVspoof 2021 challenge, in: Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 68–74.

[126] R.K. Das, Known-unknown data augmentation strategies for detection of logical access, physical access and speech deepfake attacks: ASVspoof 2021, in: Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 29–36.

[127] W. Ge, J. Patino, M. Todisco, N. Evans, Raw differentiable architecture search for speech deepfake and spoofing detection, in: Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 22–28.

[128] W.H. Kang, J. Alam, A. Fathan, CRIM's system description for the ASVSpoof2021 challenge, in: Proc. INTERSPEECH, 2021, pp. 100–106.

[129] N.M. Müller, F. Dieckmann, P. Czempin, R. Canals, K. Böttinger, J. Williams, Speech is silver, silence is golden: What do ASVspoof-trained models really learn? in: Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 55–60.

[130] H. Tak, J.-w. Jung, J. Patino, M. Kamble, M. Todisco, N. Evans, End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection, in: Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 1–8.

[131] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, G. Lavrentyeva, STC antispoofing systems for the ASVspoof2021 challenge, in: Proc. INTERSPEECH, 2021, pp. 61–67.

[132] X. Wang, X. Qin, T. Zhu, C. Wang, S. Zhang, M. Li, The DKU-CMRI system for the ASVspoof 2021 challenge: vocoder based replay channel response estimation, in: Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 16–21.

[133] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J.S. Chung, B.-J. Lee, H.-J. Yu, N. Evans, Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks, in: Proc. ICASSP, 2022, pp. 6367–6371.

[134] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, N. Evans, Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation, in: Proc. INTERSPEECH, 2022, pp. 112–119.

[135] H. Tak, M. Kamble, J. Patino, M. Todisco, N. Evans, Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing, in: Proc. ICASSP, 2022, pp. 6382–6386.

[136] R. Liu, J. Zhang, G. Gao, H. Li, Betray Oneself: A Novel Audio DeepFake Detection Model via Mono-to-Stereo Conversion, in: Proc. INTERSPEECH, 2023, pp. 3999–4003.

[137] C. Wang, J. Yi, J. Tao, C.Y. Zhang, S. Zhang, X. Chen, Detection of cross-dataset fake audio based on prosodic and pronunciation features, in: Proc. INTERSPEECH, 2023, pp. 3844–3848.

[138] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, M. Auli, Unsupervised cross-lingual representation learning for speech recognition, in: Proc. INTERSPEECH, 2021, pp. 2426–2430.

[139] W.-N. Hsu, B. Bolte, Y.-H.H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ ACM Trans. Audio Speech Lang. Process. 29 (2021) 3451–3460.

[140] X.-M. Zeng, J.-T. Zhang, K. Li, Z.-L. Liu, W.-L. Xie, Y. Song, Deepfake algorithm recognition system with augmented data for ADD 2023 challenge, in: Proc. IJCAI, 2023, pp. 31–36.

[141] Z. Teng, Q. Fu, J. White, M.E. Powell, D.C. Schmidt, SA-SASV: An end-to-end spoof-aggregated spoofing-aware speaker verification system, in: Proc. INTERSPEECH, 2022, pp. 4391–4395.

[142] J. Xue, C. Fan, J. Yi, C. Wang, Z. Wen, D. Zhang, Z. Lv, Learning from yourself: A self-distillation method for fake speech detection, in: Proc. ICASSP, 2023, pp. 1–5.

[143] Y. Xie, H. Cheng, Y. Wang, L. Ye, Learning a self-supervised domain-invariant feature representation for generalized audio deepfake detection, in: Proc. INTERSPEECH, 2023, pp. 2808–2812.

[144] Y. Yang, H. Qin, H. Zhou, C. Wang, T. Guo, K. Han, Y. Wang, A robust audio deepfake detection system via multi-view feature, in: Proc. ICASSP, 2024, pp. 13131–13135.

[145] A. Défossez, J. Copet, G. Synnaeve, Y. Adi, High fidelity neural audio compression, Trans. Mach. Learn. Res. (2023) (2023).

[146] Y.-C. Wu, I.D. Gebru, D. Marković, A. Richard, Audiodec: An open-source streaming high-fidelity neural audio codec, in: Proc. ICASSP, 2023, pp. 1–5.

[147] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, C. Feichtenhofer, Masked autoencoders that listen, Adv. Neural Inf. Process. Syst. 35 (2022) 28708–28720.

[148] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., Wavlm: Large-scale self-supervised pre-training for full stack speech processing, IEEE J. Sel. Top. Signal Process. 16 (6) (2022) 1505–1518.

[149] A. Radford, et al., Robust speech recognition via large-scale weak supervision, in: Proc. ICML, 2023, pp. 28492–28518.

[150] Y. Guo, H. Huang, X. Chen, H. Zhao, Y. Wang, Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier, in: Proc. ICASSP, 2024, pp. 12702–12706.

[151] A. Pianese, D. Cozzolino, G. Poggi, L. Verdoliva, Training-free deepfake voice recognition by leveraging large-scale pre-trained models, in: Proc. ACM Workshop on Information Hiding and Multimedia Security, 2024, pp. 289–294.

[152] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, F. Wei, BEATs: audio pre-training with acoustic tokenizers, in: Proc. ICML, 2023, pp. 5178–5193.

[153] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, S. Dubnov, Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation, in: Proc. ICASSP, 2023, pp. 1–5.

[154] A. Guzhov, F. Raue, J. Hees, A. Dengel, Audioclip: Extending clip to image, text and audio, in: Proc. ICASSP, 2022, pp. 976–980.

[155] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, E. Khoury, Generalization of audio deepfake detection, in: Odyssey, 2020, pp. 132–137.

[156] Y. Xie, H. Cheng, Y. Wang, L. Ye, Single domain generalization for audio deepfake detection., in: Proc. IJCAI, 2023, pp. 58–63.

[157] X. Chen, Y. Zhang, G. Zhu, Z. Duan, UR channel-robust synthetic speech detection system for ASVspoof 2021, in: Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 75–82.

[158] Z. Benhafid, S.A. Selouani, M.S. Yakoub, A. Amrouche, LARIHS ASSERT reassessment for logical access ASVspoof 2021 challenge, in: Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 94–99.

[159] Y. Zhang, F. Jiang, Z. Duan, One-class learning towards synthetic voice spoofing detection, IEEE Signal Process. Lett. 28 (2021) 937–941.

[160] W.H. Kang, J. Alam, A. Fathan, Investigation on activation functions for robust end-to-end spoofing attack detection system, in: Proc. INTERSPEECH, 2021, pp. 83–88.

[161] L. Zhang, X. Wang, E. Cooper, J. Yamagishi, Multi-task learning in utterance-level and segmental-level spoof detection, in: Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021.

[162] Y. Gao, T. Vuong, M. Elyasi, G. Bharaj, R. Singh, Generalized spoofing detection inspired from audio generation artifacts, in: Proc. INTERSPEECH, 2021, pp. 4184–4188.

[163] R. Yan, C. Wen, S. Zhou, T. Guo, W. Zou, X. Li, Audio deepfake detection system with neural stitching for ADD 2022, in: Proc. ICASSP, 2022, pp. 9226–9230.

[164] Y. Xie, H. Cheng, Y. Wang, L. Ye, Domain generalization via aggregation and separation for audio deepfake detection, IEEE Trans. Inf. Forensics Secur. 19 (2024) 344–358.

[165] A.K.S. Yadav, E.R. Bartusiak, K. Bhagtani, E.J. Delp, Synthetic speech attribution using self supervised audio spectrogram transformer, Electron. Imaging 35 (2023) 1–11.

[166] Y. Ren, H. Peng, L. Li, Y. Yang, Lightweight voice spoofing detection using improved one-class learning and knowledge distillation, IEEE Trans. Multimed. (2023) (2023).

[167] Y. Zhang, Z. Li, J. Lu, W. Wang, P. Zhang, Synthetic speech detection based on the temporal consistency of speaker features, IEEE Signal Process. Lett. 31 (2024) 944–948.

[168] J. Deng, Y. Ren, T. Zhang, H. Zhu, Z. Sun, VFD-net: Vocoder fingerprints detection for fake audio, in: Proc. ICASSP, 2024, pp. 12151–12155.

[169] GAN-based network decoders, 2023, https://github.com/kan-bayashi/ParallelWaveGAN.

[170] L. Cuccovillo, M. Gerhardt, P. Aichroth, Audio transformer for synthetic speech detection via formant magnitude and phase analysis, in: Proc. ICASSP, 2024, pp. 4805–4809.

[171] X. Wang, J. Yamagishi, Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end? in: Proc. ICASSP, 2024, pp. 10311–10315.

[172] H.-s. Shin, J. Heo, J.-h. Kim, C.-y. Lim, W. Kim, H.-J. Yu, HM-conformer: A conformer-based audio deepfake detection system with hierarchical pooling and multi-level classification token aggregation methods, in: Proc. ICASSP, 2024, pp. 10581–10585.

[173] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, A. Kozlov, STC antispoofing systems for the ASVspoof2019 challenge, in: Proc. INTERSPEECH, 2019, pp. 1033–1037.

[174] G. Hua, A. Teoh, H. Zhang, Towards end-to-end synthetic speech detection, IEEE Signal Process. Lett. 28 (2021) 1265–1269.

[175] X. Wang, J. Yamagishi, A comparative study on recent neural spoofing countermeasures for synthetic speech detection, in: Proc. INTERSPEECH, 2021, pp. 4259–4263.

[176] T. Chen, E. Khoury, K. Phatak, G. Sivaraman, Pindrop labs' submission to the ASVspoof 2021 challenge, in: Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 89–93.

[177] Y. Wen, Z. Lei, Y. Yang, C. Liu, M. Ma, Multi-path GMM-MobileNet based on attack algorithms and codecs for synthetic speech and deepfake detection, in: Proc. INTERSPEECH, 2022, pp. 4795–4799.

[178] I.-Y. Kwak, et al., Low-quality fake audio detection through frequency feature masking, in: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, 2022, pp. 9–17.

[179] J. Pan, S. Nie, H. Zhang, S. He, K. Zhang, S. Liang, X. Zhang, J. Tao, Speaker recognition-assisted robust audio deepfake detection, in: Proc. INTERSPEECH, 2022, pp. 4202–4206.

[180] A. Alenin, et al., A subnetwork approach for spoofing aware speaker verification, in: Proc. INTERSPEECH, 2022, pp. 2888–2892.

[181] S. Dong, J. Xue, C. Fan, K. Zhu, Y. Chen, Z. Lv, Multi-perspective information fusion Res2Net with RandomSpecmix for fake speech detection, in: Proc. IJCAI, 2023.

[182] Z. Wang, Q. Wang, J. Yao, L. Xie, The NPU-ASLP system for deepfake algorithm recognition in ADD 2023 challenge, in: Proc. IJCAI, 2023, pp. 64–69.

[183] C. Wang, J. He, J. Yi, J. Tao, C.Y. Zhang, X. Zhang, Multi-scale permutation entropy for audio deepfake detection, in: Proc. ICASSP, 2024, pp. 1406–1410.

[184] Y. Zhu, S. Koppisetti, T. Tran, G. Bharaj, SLIM: Style-linguistics mismatch model for generalized audio deepfake detection, 2024, arXiv preprint arXiv:2407.18517.

[185] H. Hu, et al., Device-robust acoustic scene classification based on two-stage categorization and data augmentation, in: Proc. DCASE, 2020.

[186] N.T. Pham, et al., Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition, Expert Syst. Appl. 230 (2023) 120608.

[187] A. Alex, L. Wang, P. Gastaldo, A. Cavallaro, Data augmentation for speech separation, Speech Commun. 152 (2023) 102949.

[188] Y. Tokozume, Y. Ushiku, T. Harada, Learning from between-class examples for deep sound recognition, in: ICLR, 2018.

[189] D.S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E.D. Cubuk, Q.V. Le, SpecAugment: A simple data augmentation method for automatic speech recognition, in: Proc. INTERSPEECH, 2019, pp. 2613–2617.

[190] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, H. Delgado, ASVspoof: the automatic speaker verification spoofing and countermeasures challenge, IEEE J. Sel. Top. Signal Process. 11 (4) (2017) 588–604.

[191] A. Babu, et al., XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale, in: Proc. INTERSPEECH, 2022, pp. 2278–2282.

[192] M. Ravanelli, Y. Bengio, Speaker recognition from raw waveform with sincnet, in: IEEE Spoken Language Technology Workshop, 2018, pp. 1021–1028.

[193] N. Zeghidour, O. Teboul, F. de Chaumont Quitry, M. Tagliasacchi, LEAF: A learnable frontend for audio classification, in: Proc. ICLR, 2021.

[194] L. Pham, P. Lam, T. Nguyen, H. Nguyen, A. Schindler, Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models, in: 2024 IEEE 5th International Symposium on the Internet of Sounds, IS2, 2024, pp. 1–5.

[195] L. Pham, D. Ngo, D. Salovic, A. Jalali, A. Schindler, P.X. Nguyen, K. Tran, H.C. Vu, Lightweight deep neural networks for acoustic scene classification and an effective visualization for presenting sound scene contexts, Appl. Acoust. 211 (2023) 109489.

[196] J. Lu, Y. Zhang, W. Wang, Z. Shang, P. Zhang, One-class knowledge distillation for spoofing speech detection, in: Proc. ICASSP, 2024, pp. 11251–11255.

[197] P. Kawa, M. Plata, P. Syga, Attack Agnostic Dataset: Towards Generalization and Stabilization of Audio DeepFake Detection, in: Proc. INTERSPEECH, 2022, pp. 4023–4027.

[198] X. Wang, J. Yamagishi, Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders, in: Proc. ICASSP, 2023, pp. 1–5.

[199] L. Pham, D. Tran, F. Skopik, A. Schindler, S. Poletti, F. David, M. Boyer, DIN-CTS: Low-complexity depthwise-inception neural network with contrastive training strategy for deepfake speech detection, 2025, arXiv preprint arXiv:2502.20225.

[200] M. Ravanelli, et al., SpeechBrain: A general-purpose speech toolkit, 2021, arXiv:2106.04624.

[201] B. Loïc, et al., Seamless: Multilingual expressive and streaming speech translation, 2023, arXiv preprint arXiv:2312.05187.

[202] A. Plaquet, H. Bredin, Powerset multi-class cross entropy loss for neural speaker diarization, in: Proc. INTERSPEECH, 2023, pp. 3222–3226.

[203] H. Bredin, pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe, in: Proc. INTERSPEECH, 2023, pp. 1983–1987.

[204] A. Chintha, et al., Recurrent convolutional structures for audio spoof and video deepfake detection, IEEE J. Sel. Top. Signal Process. 14 (5) (2020) 1024–1037.

[205] J. Deng, et al., ImageNet: A large-scale hierarchical image database, in: Proc. CVPR, 2009, pp. 248–255.

[206] T. Nguyen, F. Pernkopf, Lung sound classification using Co-tuning and stochastic normalization, IEEE Trans. Biomed. Eng. 69 (9) (2022) 2872–2882.

[207] N.M. Müller, P. Sperl, K. Böttinger, Complex-valued neural networks for voice anti-spoofing, in: Proc. INTERSPEECH, 2023, pp. 3814–3818.

[208] W. Ge, J. Patino, M. Todisco, N. Evans, Explaining deep learning models for spoofing and deepfake detection with Shapley additive explanations, in: Proc. ICASSP, 2022, pp. 6387–6391.

[209] D. Salvi, P. Bestagini, S. Tubaro, Towards frequency band explainability in synthetic speech detection, in: Proc. EUSIPCO, 2023, pp. 620–624.

[210] N. Yu, L. Chen, T. Leng, Z. Chen, X. Yi, An explainable deepfake of speech detection method with spectrograms and waveforms, J. Inf. Secur. Appl. 81 (2024) 103720.

[211] S.-Y. Lim, D.-K. Chae, S.-C. Lee, Detecting deepfake voice using explainable deep learning techniques, Appl. Sci. 12 (8) (2022) 3926.

[212] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proc. International Conference on Neural Information Processing Systems, 2017, pp. 4768–4777.

[213] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[214] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proc. ICLR, 2015.

[215] D. Smilkov, N. Thorat, B. Kim, F.B. Viégas, M. Wattenberg, SmoothGrad: removing noise by adding noise, 2017, CoRR abs/1706.03825.

[216] R.A. Gougeh, Z. Nu, Z. Zilic, Optimizing Auditory Immersion Safety on Edge Devices: An On-Device Sound Event Detection System, in: Proc. the Speaker and Language Recognition Workshop, 2024, pp. 225–231.

[217] J.J. Bird, A. Lotfi, Real-time detection of AI-generated speech for DeepFake voice conversion, 2023, arXiv preprint arXiv:2308.12734.

[218] J.J. Mathew, R. Ahsan, S. Furukawa, J.G.K. Kumar, H. Pallan, A.S. Padda, S. Adamski, M. Reddiboina, A. Pankajakshan, Towards the development of a real-time deepfake audio detection system in communication platforms, 2024, arXiv preprint arXiv:2403.11778.

[219] Q. Zhang, S. Wen, T. Hu, Audio deepfake detection with self-supervised xls-r and sls classifier, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 6765–6773.

[220] Y. Guo, H. Huang, X. Chen, H. Zhao, Y. Wang, Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier, in: Proc. ICASSP, 2024, pp. 12702–12706.

[221] X. Zhang, J. Yi, C. Wang, C.Y. Zhang, S. Zeng, J. Tao, What to remember: Self-adaptive continual learning for audio deepfake detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, (no. 17) 2024, pp. 19569–19577.

[222] Y. Chen, J. Yi, C. Fan, J. Tao, Y. Ren, S. Zeng, C.Y. Zhang, X. Yan, H. Gu, J. Xue, et al., Region-based optimization in continual learning for audio deepfake detection, 2024, arXiv preprint arXiv:2412.11551.

[223] S. Hou, Y. Ju, C. Sun, S. Jia, L. Ke, R. Zhou, A. Nikolich, S. Lyu, DeepFake-O-Meter v2. 0: An open platform for DeepFake detection, 2024, arXiv preprint arXiv:2404.13146.