Benchmarking Audio Deepfake Detection Robustness in Real-world Communication Scenarios

Haohan Shi[1], Xiyu Shi[1], Safak Dogan[1], Saif Alzubi[2], Tianjin Huang[2], Yunxiao Zhang[2]
[1]Institute for Digital Technologies, Loughborough University London, London E20 3BS, UK
[2]Department of Computer Science, University of Exeter, Exeter EX4 4QE, UK
{h.shi, x.shi, s.dogan}@lboro.ac.uk, {s.m.y.alzubi, t.huang2, y.zhang12}@exeter.ac.uk

***Abstract*—Existing Audio Deepfake Detection (ADD) systems often struggle to generalise effectively due to the significantly degraded audio quality caused by audio codec compression and channel transmission effects in real-world communication scenarios. To address this challenge, we developed a rigorous benchmark to evaluate the performance of the ADD system under such scenarios. We introduced ADD-C, a new test dataset to evaluate the robustness of ADD systems under diverse communication conditions, including different combinations of audio codecs for compression and packet loss rates. Benchmarking three baseline ADD models on the ADD-C dataset demonstrated a significant decline in robustness under such conditions. A novel Data Augmentation (DA) strategy was proposed to improve the robustness of ADD systems. Experimental results demonstrated that the proposed approach significantly enhances the performance of ADD systems on the proposed ADD-C dataset. Our benchmark can assist future efforts towards building practical and robustly generalisable ADD systems.**

***Index Terms*—Audio Deepfake Detection, Audio Signal Processing, Audio Codec, Robustness, Wireless Communication.**

I. Introduction

Recent advancements in AI-based Text-to-Speech (TTS) and Voice Conversion (VC) technology make it easier to synthesise natural, human-like speech from text or audio inputs [1]. Such technology significantly enhances the convenience in various aspects of our daily lives, e.g., e-book readers, voice assistants, and smart home devices. However, the misuse for malicious purposes poses emerging threats and challenges to security [2]. In 2020, fraudsters used AI-generated deepfake audio to impersonate a company's director, deceiving a branch manager into transferring $35 million [3].

In response to such attacks, Audio Deepfake Detection (ADD) aims to identify AI-generated synthesised audio to determine its authenticity. Recent advancements, such as the Audio Speaker Verification (ASV) spoof challenge series [4], have significantly contributed to the progress of ADD by providing standardised benchmarks and encouraging the development of detection models. These efforts have led to notable improvements in detecting deepfake audio across various generation techniques.

However, existing methods are trained on clean, high-quality audio and often fail to generalise well to real-world communication scenarios, where audio codec compression and channel transmission effects degrade audio quality [5]. These challenges are particularly evident in wireless communications based on Voice over Long Term Evolution (VoLTE) [6] and Voice over Internet Protocol (VoIP) systems [7]. The lack of robustness highlights the need for more practical ADD systems capable of effectively detecting deepfake audio in real-world communication scenarios.

Therefore, this paper focuses on addressing the performance degradation of ADD systems caused by audio codec compression and channel transmission effects in real-world communication scenarios. By simulating the wireless communication environments, we systematically analyse and improve the robustness of ADD systems. Our contributions are as follows:

- To the best of our knowledge, this is the first study to systematically investigate the impact and robustness of real-world communication scenarios on ADD systems.
- A new benchmarking framework is designed to systematically train and evaluate the robustness of ADD systems under various communication conditions.
- We propose a new test dataset, ADD-C, to assess the performance of ADD systems in real-world communication scenarios. ADD-C includes six evaluation conditions: one clean condition and five real-world communication conditions. Each real-world condition includes simulating audio codec compression using six widely used speech codecs in VoLTE and VoIP communication systems, as well as simulating channel transmission effects under five different Packet Loss Rates (PLRs). Benchmarking three baseline ADD models reveals a significant decline in performance, highlighting the need for improving the robustness of ADDs.
- A novel Data Augmentation (DA) strategy is proposed to address the weak robustness of ADD systems. Experimental results demonstrate that our approach significantly enhances the robustness of ADD systems in real-world communication scenarios.

II. Related Work

The ADD task has gained increasing importance over time, leading to the development of various methodologies. These methods can be broadly categorised into machine learning and deep learning-based approaches.

This research was funded by Loughborough University (Grant No. GS1016) and the China Scholarship Council (Grant No. 202208060237).

Traditional machine learning-based approaches rely on handcrafted acoustic features. For example, Mel-frequency Cepstral Coefficients (MFCC) have been widely used in classifiers such as support vector machines, AdaBoost, decision trees, etc., demonstrating their effectiveness in ADD tasks [8]. Additionally, other acoustic features, including Constant-Q Cepstral Coefficients (CQCC), Linear-Frequency Cepstral Coefficients (LFCC), Mel-spectrograms, and constant-Q-transform [9]–[11], have also been extensively utilised.

With the development of deep learning, competitions such as ASVspoof [4] have emerged, leading to a more diverse range of detection and feature extraction methods. Convolutional Neural Networks (CNN) [12], Long Short-Term Memory (LSTM) networks [13], and attention mechanisms [14] have been widely used to enhance detection accuracy and feature extraction efficiency. To further improve model performance, [15] proposed a CNN-LSTM-based model that combines MFCC, Mel spectrogram, CQCC, and CQT features to tackle the ADD task. Similarly, [16] introduced Res-TSSDNet, which utilises the fusion of raw waveform and spectrogram representations to achieve better detection accuracy.

Some studies have explored the use of perceptual features for ADD. For instance, [17] utilises frequency band information and complementary real-imaginary spectrogram features to address ADD challenges, while [18] applies a self-attention mechanism to extract phoneme-based representations. Additionally, physiological characteristics such as breathing patterns [19] and human vocal tract features [20] have been investigated to provide deeper insights.

With the advancement of deep learning, end-to-end ADD methods have become popular, enabling automatic feature extraction without manual design. Some methods utilise pre-trained self-supervised models to extract features, such as Wav2Vec [21], [22], WavLM [23], and XLS-R [9]. Additionally, [24] modified the original RawNet2 architecture to classify audio authenticity from raw waveform inputs directly.

Despite these advancements, ensuring the robustness of ADD systems remains a significant challenge. To enhance model generalisation to unseen synthesis techniques and real-world recording, [25] proposed an aggregation and separation domain generalisation network, which integrates adversarial training and domain adaptation. Similarly, [26] introduced the GMM-MobileNet model, which employs a multi-path structure to enhance accuracy in unseen deepfake algorithms.

However, a critical research gap remains: the impact of real-world communication on ADD has been largely overlooked. In such communication scenarios, the primary considerations encompass both channel transmission and audio codec [27]. Within wireless channels, impairments such as bandwidth mismatch, latency, jitter, and PLR can introduce distorted transmissions, while audio codecs often induce compression artifacts. Prior studies indicate that channel-induced data loss degrades the performance of audio-based feature systems [28], while codec-induced compression artifacts result in diminished audio quality as high-frequency information is lost [29], which reduces the robustness of the ADD system. Additionally, over mobile or internet networks, the combined effects of audio codec compression and channel transmission degradation reduce the speech quality [28]. Such ADD tasks in real-world communication scenarios have not been systematically studied, highlighting the importance of our research.

III. Benchmark Design and Baseline Evaluation

A. Real-world Communication Simulation

Six speech codecs were selected to simulate real-world communication scenarios: AMR-WB [30], EVS [31], IVAS [32], OPUS [33], Speex (WB) [34], and SILK [35]. These codecs were selected for their diversity and broad applicability, which span various use cases ranging from cellular network voice calls to VoIP. This selection ensures that the simulated experimental environment closely approximates real-world communication scenarios. Details of the codecs are presented in TABLE I.

TABLE I
Details of the Selected Codec

Index	Codec	Sample Rate(kHz)	Bitrate(kbps)
1	AMR-WB	16	6.60-23.85
2	EVS	8,16,32,48	5.90-128
3	IVAS	8,16,32,48	13.20-512
4	OPUS	8-48	6-510
5	Speex(WB)	8,16,32	2-44
6	SILK	8-24	6-40

Five PLRs, 0%, 1%, 5%, 10%, and 20%, were selected for this study to realistically simulate the impact of network congestion, wireless interference, and other transmission impairments. These rates represent various communication environments, ranging from ideal transmission conditions to severely degraded channels with higher PLRs. This systematic approach enables a comprehensive analysis of the effects of varying communication conditions on the ADD system.

B. ADD-C Dataset Building

To systematically evaluate the impact of real-world communication scenarios on ADD systems, we propose the ADD-C test dataset, which is based on six publicly available speech datasets, Fake-or-Real (FoR) [37], Wavefake [38], LJSpeech [39], MLAAD [40], M-AILABS [41] and ASVspoof2021 Logical Access (ASV) [4]. The details of these datasets, including the number of real and fake utterances, are listed in TABLE II.

TABLE II
Details of the Selected Datasets

Dataset	Real	Fake	Language	Algorithms
FoR	34605	34695	English	7
Wavefake & LJSpeech (W&L)	13100	91700	English	7
MLAAD & M-AILABS (M&M)	69853	5000	English	5
ASV	12483	108978	English	17
Total	130041	240373	-	36

The Wavefake and MLAAD datasets contain fake utterances generated using LJSpeech and M-AILABS as source data, respectively, while LJSpeech and M-AILABS consist of real human speech recordings. These four datasets are used in pairs, denoted as W&L and M&M. In total, the datasets contain 130,041 real and 240,373 fake utterances, involving 36 types of deepfake algorithms. To ensure consistency, all

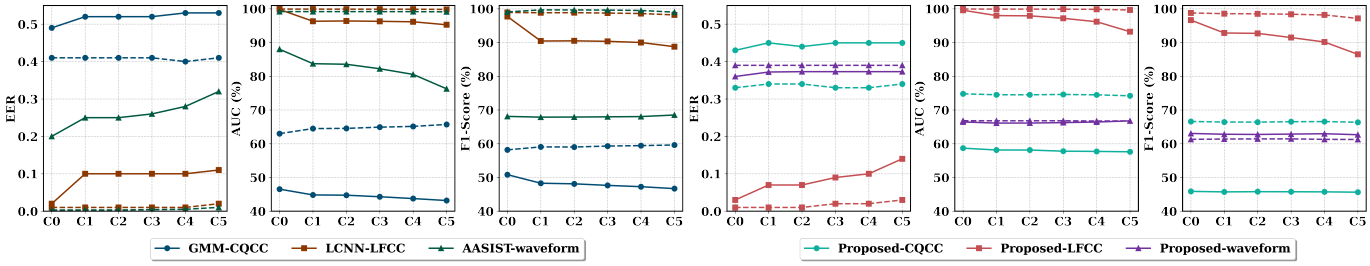

Fig. 1. Results of EER, AUC and F1-score on ADD-C test dataset. The first three subfigures represent the baseline models GMM [4], LCNN [4], and AASIST [36]. The last three subfigures represent the proposed models. Solid and dashed lines denote training on the Original and Augmented dataset, respectively.

four datasets are converted to a single-channel 16-bit Pulse-Code Modulation format with a sampling rate of 16kHz.

The ADD-C test dataset consists of six conditions ($C_0 - C_5$). C_0 represents the clean condition and was built with 500 real and 500 fake utterances selected from four datasets in TABLE II, respectively, without any codec compression and transmission effects. C_1-C_5 represent five distinct communication conditions defined as follows:

$$C_n = \sum_{i=1}^{6} T\left(C_0, \text{Codec}_i, \text{PLR}_n\right), n = 1, \ldots, 5, \quad (1)$$

where Codec_i and PLR_n represent six codecs in TABLE I and the five PLRs (i.e., 0%, 1%, 5%, 10%, and 20%), respectively. $T(\cdot)$ is the operation performed on C_0, simulating six codecs' compression under one of the five PLRs. All selected utterances were non-overlap and removed from the source dataset. Details are shown in TABLE III, the proportion of real and fake under each condition is equal.

TABLE III
THE PROPOSED ADD-C TEST DATASET

Condition	C_0	C_1	C_2	C_3	C_4	C_5
PLR(%)	-	0	1	5	10	20
Total utterances	4000	24000	24000	24000	24000	24000

C. Evaluation and Results of Baseline Models

To systematically assess the impact of real-world communication scenarios on ADD systems, particularly considering audio codec compression and channel transmission effects, three baseline ADD models were selected for evaluation: GMM [4], LCNN [4], and AASIST [36], utilising CQCC, LFCC, and raw waveform as acoustic features, respectively. For a fair comparison, the four datasets in TABLE II were merged to form a unified Original dataset, with a split ratio of 80%:20% for training and validation. All baseline models used the hyperparameters specified in the referenced literature.

The evaluation was conducted on the ADD-C test dataset. Three evaluation metrics, Equal Error Rate (EER), Area Under the Curve (AUC), and F1-score were selected to assess the models' robustness and performance. EER refers to the error rate in binary classification systems when the false positive rate equals the false negative rate. A lower EER indicates better overall model performance, while a higher AUC and F1-score represent better discrimination performance.

The results of baseline models trained on the Original dataset are shown with solid lines in the first three subfigures of Fig. 1. A notable performance drop is observed from C_0 to C_1-C_5. Specifically, when comparing C_0 to C_1, the baseline models experienced an average degradation of 5.30% in EER, 3.16% in AUC and 3.34% in F1-score. While most metrics exhibit a consistent decline from C_1 to C_5, a slight increase (<1%) in the F1-score of AASIST was observed. This minor fluctuation was likely attributed to experimental noise rather than a fundamental improvement in robustness. Overall, these results highlight the substantial impact of various communication conditions on ADD robustness, indicating that the system's ability to distinguish between real and fake audio significantly declines in real-world communication scenarios.

IV. PROPOSED METHOD

In this section, three models and a DA strategy are proposed to address the performance decline of ADD systems caused by real-world communication scenarios.

A. Model Architecture

Inspired by the ADD frameworks presented in [4], [26], three models were designed to handle different input feature representations. Each model shares a typical architecture comprising two main components: feature extraction and classification. The proposed architectures are shown in Fig. 2.

All audios were cut to 4s prior to feature extraction, and zero-padding was applied to audio clips shorter than 4s. Let $x = [x(1), x(2), \cdots, x(L)]^T \in \mathbb{R}$ denote the original speech signal in time-domain, where $x(L)$ corresponds to the L-th sample point and L equal to 64,000. Let us define the input into the proposed model as $X \in \mathbb{R}^{B \times N \times D \times T}$, where B is the batch size, N is the number of channels, D is the dimension of the feature, and T is the number of time frame. Due to the difference of input acoustic features, there are three types of X before sending into the feature extractor, labelled as $X_{lfcc} \in \mathbb{R}^{B \times N \times D_1 \times T_1}$, $X_{cqcc} \in \mathbb{R}^{B \times N \times D_2 \times T_2}$ and $X_{wav} \in \mathbb{R}^{B \times N \times D_3 \times T_3}$, as shown in Fig. 2 (a). All audio inputs in this study are single-channel (mono) signals, resulting in $N = 1$. Both LFCC and CQCC are 2D time-frequency representations, with each feature map having a dimensionality of 60, which consists of 20-dimensional static feature coefficients, 20-dimensional first-order delta coefficients, and 20-dimensional second-order delta coefficients, leading to $D_1 = D_2 = 60$. In contrast, the raw waveform is a 1D feature representation, and the input shape corresponds to the number of sampled points, resulting in $D_3 = 1$ and $T_3 = L$. The default hop lengths used during feature calculation for LFCC and CQCC in the baseline

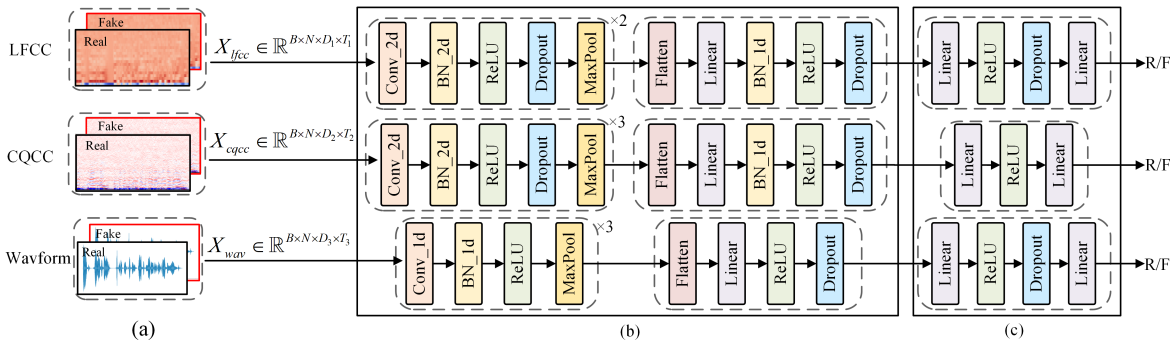

Fig. 2. Architectures of the proposed models. (a) Different inputs of acoustic features; (b) Feature extractor; (c) Classifier.

models are 512 and 128, respectively, resulting in $T_1 = 126$ and $T_2 = 501$. Finally, the output of the feature extractor passes through the classifier and outputs the authenticity of the input signal.

B. Data Augmentation (DA) Strategy

Fig. 3 shows the proposed novel DA strategy for mitigating performance degradation in ADD systems.

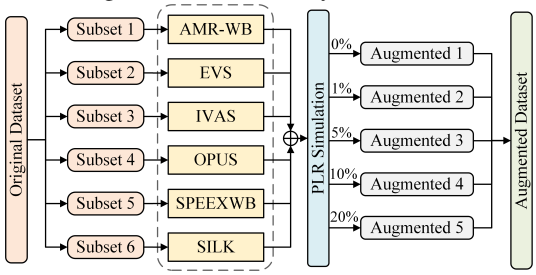

Fig. 3. Proposed DA strategy

The Original dataset was first partitioned into six equal subsets to ensure balanced representation of deepfake algorithms and speaker distributions, maintaining overall data diversity. Each subset was then processed using a speech codec to simulate codec compression, followed by a PLR simulator to simulate channel transmission effects. This produced six augmented datasets, each corresponding to a different PLR. All datasets generated under the five PLRs were then merged to form the final Augmented dataset, which is five times that of the Original dataset, substantially enriching the training corpus and enhancing model generalisation in real-world communication scenarios.

V. Experiments and Results

A. Training Setup

The Augmented dataset is constructed by applying the proposed DA strategy to the Original dataset. It comprises a total of 1,832,070 utterances, including 640,205 real and 1,191,865 fake samples. For model training, the Augmented dataset is split into 80% for training and 20% for validation. Models are trained for five epochs with a batchsize of 256 using the Adam optimiser [42]. Early Stopping [43] with a patience of three is employed to prevent overfitting. The models are trained using Cross-Entropy Loss:

$$\mathcal{L} = -\frac{1}{J}\sum_{j=1}^{J}\left[y_j \cdot \log(\hat{y}_j) + (1-y_j)\cdot\log(1-\hat{y}_j)\right], \quad (2)$$

where J represents the total number of samples, y_j the binary ground-truth label (0 or 1), and $\hat{y}_j$ the predicted probability.

B. Results and Discussion

To evaluate the effectiveness of the proposed DA strategy while ensuring a fair comparison, a two-stage evaluation was employed. First, extending the results of the baseline models trained on the Original dataset presented in Section III-C, the three baseline models were retrained using the Augmented dataset. Their performance on the ADD-C test dataset is illustrated by the dashed lines in the first three subfigures of Fig. 1. Second, the proposed models were trained separately on the Original and Augmented datasets. The corresponding results are shown in the solid and dashed lines of the last three subfigures in Fig. 1, respectively.

Compared to the notable performance decline and limited robustness shown by the baseline models trained on the Original dataset, those trained on the Augmented dataset demonstrate significant performance improvements, exhibiting strong stability without noticeable fluctuations or degradation across varying communication conditions.

For the proposed models trained on the Original dataset, a clear performance drop is observed from C_0 to C_1-C_5. Specifically, comparing C_0 to C_1 yields an average degradation of 2.33% (EER), 0.49% (AUC), and 1.40% (F1-score). Apart from a negligible fluctuation in the AUC of the proposed waveform-based model, all metrics show a consistent downward trend from C_1 to C_5, which aligns with the results in Section III-C. In contrast, the proposed models trained on the Augmented dataset exhibit stable performance across all metrics. The EER remains unchanged, with a slight improvement of 0.003% from C_0 to C_1. AUC and F1-score also remain highly stable, with a negligible decrease of 0.1%.

Notably, unlike the models trained on the Original dataset that suffer from a significant performance degradation, those trained on the Augmented dataset exhibit consistent stability across all evaluation metrics from C_0 to C_5, without noticeable fluctuations or drops. These results demonstrate that the proposed DA strategy effectively enhances the dataset by introducing diverse variations that simulate channel transmission and codec compression, thereby improving the model's generalisation under real-world communication scenarios.

In conclusion, real-world communication scenarios significantly impact the robustness of ADD systems, while the

proposed DA strategy can successfully mitigate the degradation caused by codec compression and channel transmission distortions, enhancing the robustness and ensuring more reliable ADD system deployment in realistic and practical communication environments.

VI. Conclusion

This work systematically investigates the impact of real-world communication scenarios on ADD systems. A new benchmark was established to assess the robustness of ADD systems under various communication conditions, accompanied by introducing a new test dataset, ADD-C. Furthermore, a novel DA strategy was proposed to effectively mitigate the degradation of robustness in various communication conditions. The proposed benchmark and methodology lay a solid foundation for future research to develop more robust and security-critical ADD systems.

References

[1] C. Bisogni, V. Loia, M. Nappi, and C. Pero, "Acoustic features analysis for explainable machine learning-based audio spoofing detection," *Computer Vision and Image Understanding*, vol. 249, p. 104145, 2024.

[2] M. Rabhi, S. Bakiras, and R. Di Pietro, "Audio-deepfake detection: Adversarial attacks and countermeasures," *Expert Systems with Applications*, vol. 250, p. 123941, 2024.

[3] T. Brewster, "Fraudsters cloned company director's voice in $35 million heist, police find," *Forbes*, 2021.

[4] X. Liu *et al.*, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[5] A. Cohen *et al.*, "A study on data augmentation in voice anti-spoofing," *Speech Communication*, vol. 141, pp. 56–67, 2022.

[6] S. Sesia, I. Toufik, and M. Baker, *LTE-the UMTS long term evolution: from theory to practice*. Wiley, 2011.

[7] B. Goode, "Voice over internet protocol (voip)," *Proceedings of the IEEE*, vol. 90, no. 9, pp. 1495–1517, 2002.

[8] A. Hamza *et al.*, "Deepfake audio detection via mfcc features using machine learning," *IEEE Access*, vol. 10, pp. 134 018–134 028, 2022.

[9] Q. Zhang *et al.*, "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.

[10] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "A comparative study on physical and perceptual features for deepfake audio detection," in *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 2022, pp. 35–41.

[11] N. Chakravarty and M. Dua, "A lightweight feature extraction technique for deepfake audio detection," *Multimedia Tools and Applications*, vol. 83, no. 26, pp. 67 443–67 467, 2024.

[12] S. M *et al.*, "Classification of deep fake audio using mfcc technique," in *IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems*, 2024, pp. 1–6.

[13] T. M. Wani *et al.*, "Detecting audio deepfakes: Integrating cnn and bilstm with multi-feature concatenation," in *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, 2024, pp. 271–276.

[14] T. Kanwal *et al.*, "Fake speech detection using vggish with attention block," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 35, 2024.

[15] T. M. Wani *et al.*, "Detecting audio deepfakes: Integrating cnn and bilstm with multi-feature concatenation," in *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, 2024, pp. 271–276.

[16] N. Yu, L. Chen, T. Leng, Z. Chen, and X. Yi, "An explainable deepfake of speech detection method with spectrograms and waveforms," *Journal of Information Security and Applications*, vol. 81, p. 103720, 2024.

[17] J. Xue *et al.*, "Audio deepfake detection based on a combination of f0 information and real plus imaginary spectrogram features," in *Proceedings of the 1st international workshop on deepfake detection for audio multimedia*, 2022, pp. 19–26.

[18] H. Dhamyal, A. Ali, I. A. Qazi, and A. A. Raza, "Using self attention dnns to discover phonemic features for audio deep fake detection," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 1178–1184.

[19] T.-P. Doan, L. Nguyen-Vu, S. Jung, and K. Hong, "Bts-e: Audio deepfake detection using breathing-talking-silence encoder," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[20] L. Blue *et al.*, "Who are you(i really wanna know)? detecting audio deepfakes through vocal tract reconstruction," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 2691–2708.

[21] J. M. Martín-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 9241–9245.

[22] H. Tak *et al.*, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," *arXiv preprint arXiv:2202.12233*, 2022.

[23] Y. Guo *et al.*, "Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 12 702–12 706.

[24] H. Tak *et al.*, "End-to-end anti-spoofing with rawnet2," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 6369–6373.

[25] Y. Xie *et al.*, "Domain generalization via aggregation and separation for audio deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 344–358, 2024.

[26] Y. Wen, Z. Lei, Y. Yang, C. Liu, and M. Ma, "Multi-path gmm-mobilenet based on attack algorithms and codecs for synthetic speech and deepfake detection." in *INTERSPEECH*, 2022, pp. 4795–4799.

[27] A. F. Molisch, *Wireless communications*. John Wiley & Sons, 2012, vol. 34.

[28] L. Besacier *et al.*, "Overview of compression and packet loss effects in speech biometrics," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 150, no. 6, pp. 372–376, 2003.

[29] M. Todisco, H. Delgado, and N. Evans, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[30] B. Bessette *et al.*, "The adaptive multirate wideband speech codec (amr-wb)," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 620–636, 2002.

[31] S. Bruhn *et al.*, "Standardization of the new 3gpp evs codec," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5703–5707.

[32] ETSI, "Lte; 5g; codec for immersive voice and audio services - detailed algorithmic description incl. rtp payload format and sdp parameter definitions," 2024. [Online]. Available: https://www.etsi.org/

[33] J.-M. Valin *et al.*, "High-quality, low-delay music coding in the opus codec," *arXiv preprint arXiv:1602.04845*, 2016.

[34] J.-M. Valin, "Speex: A free codec for free speech," *arXiv preprint arXiv:1602.08668*, 2016.

[35] H. Astrom *et al.*, "Rtp payload format and file storage format for silk speech and audio codec," 2009. [Online]. Available: https://datatracker.ietf.org/doc/draft-spittka-silk-payload-format/00/

[36] J.-w. Jung *et al.*, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 6367–6371.

[37] R. Reimao and V. Tzerpos, "For: A dataset for synthetic speech detection," in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2019, pp. 1–10.

[38] J. Frank and L. Schönherr, "Wavefake: A data set to facilitate audio deepfake detection," *arXiv preprint arXiv:2111.02813*, 2021.

[39] K. Ito and L. Johnson, "The lj speech dataset," 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[40] N. M. Müller *et al.*, "Mlaad: The multi-language audio anti-spoofing dataset," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–7.

[41] "The m-ailabs speech dataset," 2024. [Online]. Available: https://github.com/imdatceleste/m-ailabs-dataset

[42] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[43] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM review*, vol. 60, no. 2, pp. 223–311, 2018.