Lightweight MFCC-EffNet for Spoofing Detection in Low-Resource South Asian Languages

Muhammad Hamza[1]
mhamza2k24@gmail.com
Junaid Mir[1]
junaid.mir@uettaxila.edu.pk
Hafsa Ilyas[2]
hafsailyas13@gmail.com
Muhammad Haroon Yousaf[3]
haroon.yousaf@uettaxila.edu.pk
Ali Javed[2]
ali.javed@uettaxila.edu.pk
Ahmed Zoha[4]
ahmed.zoha@glasgow.ac.uk

[1] Department of Electrical Engineering,
University of Engineering and
Technology Taxila, Pakistan
[2] Department of Software Engineering,
University of Engineering and
Technology Taxila, Pakistan
[3] Department of Computer Engineering,
University of Engineering and
Technology Taxila, Pakistan
[4] James Watt School of Engineering,
University of Glasgow, United Kingdom

Abstract

Deepfakes threaten the reliability of speech technologies, yet most anti-spoofing approaches are designed for high-resource languages, leaving low-resource and multilingual scenarios underexplored. This challenge is particularly acute in South Asia, where code-switching and diverse acoustic conditions complicate the detection process. To address this, we introduce a spoofed speech dataset comprising bona fide and manipulated utterances in Urdu and Hindi, with natural code-switching into English, enabling benchmarking in multilingual and low-resource settings. We further propose MFCC-EffNet, a lightweight spoofing detection framework that fuses MFCC and spectrogram features through a modified EfficientNetV2 with cross-attention. Evaluations across multiple datasets show strong generalization, with only 250k parameters and low inference latency, making it well-suited for real-time and edge deployment. Our contributions lay the groundwork for advancing robust and efficient anti-spoofing in multilingual contexts.

1 Introduction

Automatic speaker verification (ASV) systems have become integral to biometric security, enabling seamless authentication in mobile devices, smart assistants, banking platforms, and forensic investigations [8, 17]. By leveraging unique vocal traits, these systems provide both convenience and protection. However, their reliability is increasingly threatened by presentation attacks (spoofing), in which adversaries employ manipulated or artificially generated speech to deceive the system [33]. Spoofing methods include replay attacks (playing back genuine recordings), voice conversion (mimicking a target speaker's characteristics), and text-to-speech (TTS) or deepfake audio synthesis using advanced generative AI models [17]. These threats pose serious risks ranging from financial fraud to misinformation campaigns, highlighting the urgent need for robust anti-spoofing countermeasures.

Recent advancements in speech synthesis have further escalated this challenge. Modern TTS systems use neural vocoders to generate highly natural speech from text, while voice conversion (VC) techniques transfer prosodic and spectral features across speakers [20]. Such advances have enabled the creation of deepfake audio that is often indistinguishable from genuine speech, making detection increasingly difficult for both humans and automated systems. The ASVspoof Challenge series [32, 34, 38] has been instrumental in driving progress by releasing standardized datasets and evaluation protocols. However, these corpora remain primarily focused on English and a few high-resource languages, leaving significant gaps in multilingual and low-resource contexts.

Over half a billion people in South Asia speak Urdu and Hindi, where natural code-switching with English is a prominent feature of everyday communication. However, existing spoofed speech datasets fail to capture these multilingual conversational dynamics, leaving the robustness of anti-spoofing systems for this population largely unexplored. To bridge this gap, we present a spoofed speech dataset and a detection framework in this work. The primary contributions are

- A spoofed speech dataset in Urdu and Hindi is presented that incorporates natural code-switching with English. The dataset features conversational styles from bilingual Urdu and Hindi speakers, laying the groundwork for developing generalizable anti-spoofing solutions in low-resource, code-switching contexts.

- We propose a spoofed speech detection framework, MFCC-EffNet, which fuses MFCC [39] and spectrogram features (extracted through a truncated EfficientNetV2 [30] backbone) via cross-attention, enabling dual-branch modeling of phonetic cues and spectral artifacts.

- MFCC-EffNet employs a lightweight Fused-MBConv backbone (250k parameters) and language-agnostic training with adversarial augmentation, achieving $< 1\%$ EER across the proposed dataset.

The rest of the paper is structured as follows: Section 2 presents the related literature, Section 3 describes the proposed methodology, Section 4 discusses the experimental setup and results, and Section 5 concludes the paper.

2 Literature Review

Audio spoofing detection has evolved through the ASVspoof challenges [1], with a focus on feature representation, computational efficiency, and cross-lingual robustness. Starting with ASVspoof 2015 [37], subsequent editions such as ASVspoof 2019 [32], ASVspoof 2021 [38], and ASVspoof-5 [34] have provided large-scale benchmark datasets and evaluation protocols. These corpora include both Logical Access (LA) tasks, targeting synthetic and converted speech, and Physical Access (PA) tasks, addressing replayed audio. Notably, ASVspoof 2021 introduced a deepfake detection track to tackle neural speech synthesis. Other corpora such as FMFCC-a (Chinese) [39], HABLA (Spanish) [9], CFAD (Chinese)

[23], and MLAAD (38 languages) [26] have further expanded the scope of spoof detection. However, all these datasets lack code-switching dynamics and are dominated by English or other high-resource languages.

Parallel to the development of datasets, researchers have proposed diverse architectures for spoof detection. Early handcrafted features like MFCCs and CQCCs [31] captured phonetic and spectral cues but struggled with neural vocoder outputs. Deep learning shifted the paradigm to end-to-end systems, particularly 1D CNNs on raw waveforms captured fine-grained temporal anomalies [8]. RawNet [16] and A-RawNet2 [14] achieved 4.61% EER on ASVspoof 2019 LA [32] through residual blocks and attention mechanisms. RawNet3 [36] added dynamic gradient masking for robustness but increased the parameters to 5.2M, limiting mobile deployment. These models often miss spectral artifacts detectable in time-frequency representations [33], addressed by 2D CNNs like DualSpecNet [12], which reduced EER to 2.2% on ASVspoof 2021 DF but required $\sim$ 10M parameters. Hybrid fusion models that combine MFCCs, spectrograms, and raw audio inputs achieve competitive performance by leveraging complementary representations [17]. More recently, efficient CNN-based models such as BC-ResMax [6] (0.47% EER, 0.8M parameters) and DDWS-Conv [18] have shown strong efficiency. Similarly, self-attention based fusion systems [15] achieved a 74.6% EER reduction but relied on computationally heavy attention layers (3.5M parameters), making them less practical for edge deployment.

Self-supervised learning (SSL) has also gained attraction. Wav2vec 2.0 [24] improved cross-lingual robustness with a 12% relative EER reduction, while AASIST [35] and its Urdu-focused extension AASIST-L [27] demonstrated strong deepfake detection (0.52% EER). However, SSL-based systems often require large-scale pretraining and significant computational resources. Cross-lingual evaluations further highlight limitations: conventional detectors tend to degrade sharply, while linguistically adaptive models, such as SpeakerNet [13], achieve as low as 0.02% EER on English–Urdu mixtures. Indo-Aryan languages also exhibit distinct high-frequency spoofing artifacts [4], emphasizing the need for multilingual adaptation and generalization. Transformer-based spectrogram models (AST [10], SSAST [11], HTS-AT [5]) achieved EERs in the range of 2.85%–2.52%, but their large parameter counts (30–85M) make them computationally expensive, limiting deployment on edge devices.

Despite these advances, a fundamental challenge persists: the absence of large-scale datasets for low-resource and multilingual languages. Existing corpora focus mainly on English and a few high-resource languages, leaving multilingual and code-switched scenarios underrepresented. This limitation restricts model generalization across unseen spoofing techniques, languages, and recording conditions, underscoring the need for datasets and lightweight models tailored to low-resource multilingual contexts. Spectrogram-only CNNs, such as EfficientNetV2, offer faster inference but lack phonetic granularity, which is crucial for languages with rich vocalic inventories, such as Hindi and Urdu. These trade-offs between accuracy and efficiency are particularly acute for South Asian low-resource languages, where both phoneme-level cues (captured by MFCCs) and spectral anomalies (captured by Mel spectrograms) are needed to detect deepfakes.

Datasets	Split	# Bona fide	# Spoofed	# Male	# Female
Proposed (Urdu)	Train	3,418	15,300	11	7
	Dev	3,857	13,600	11	6
	Eval	5,431	59,500	21	14
Proposed (Hindi)	Train	2,287	20,500	14	7
	Dev	2,520	21,000	14	7
	Eval	3,329	85,895	28	14
Urdu-DF	Train	2,199	2,199	7	4
	Dev	599	597	1	2
	Eval	600	600	2	1
ASVspoof 2019	Train	2,580	22,800	11	8
	Dev	2,487	22,296	13	7
	Eval	7,355	63,882	24	14
ASVspoof 2021	Eval	7,558	145,638	24	14

Table 1: Dataset Statistics.

3 Methodology

3.1 Datasets

This work employs multiple datasets, detailed below, for robust evaluation of the proposed MFCC-Effnet anti-spoofing framework under diverse and challenging conditions.

3.1.1 Proposed Dataset

The dataset proposed in this work is the first large-scale multilingual spoofing corpus designed for low-resource South Asian languages. It includes both bona fide and spoofed speech in Urdu and Hindi with natural code-switching into English, enabling robust evaluation in linguistically diverse scenarios. The bona fide samples are sourced from the MAV-Celeb dataset [29], which contains speech from 154 celebrity speakers recorded under realistic conditions with natural background noise, language switching, and multi-speaker conversational contexts. To generate spoofed utterances while preserving the multilingual and multi-speaker characteristics of bona fide speech, we employ five state-of-the-art VC systems: FreeVC [21], Diff-Hier-VC [7], HierSpeech++ [19], KNN-VC [3], and SeedVC [22]. For each target speaker, one utterance is selected as the reference, and spoofed samples are created by converting source utterances from other speakers into the target voice. This process yields ten spoofed audios per speaker per VC model. All audio files are stored in FLAC format with a 16 kHz sampling rate and an average utterance duration of 15 seconds. In total, the dataset contains 215,795 spoofed and 20,842 bona fide utterances from 154 speakers (70 Urdu and 84 Hindi). The corpus is partitioned into training, development, and evaluation subsets in a 25:25:50 ratio, balanced across both gender and language. This split follows the design philosophy of the ASVspoof challenges, where the evaluation partition is deliberately larger than training and development to encourage generalization. Table 1 summarizes the language-wise statistics and the train–dev–eval distribution of the proposed dataset.

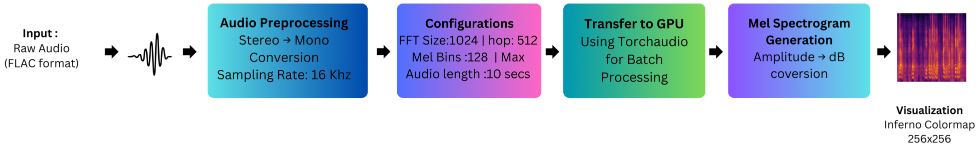

Figure 1: Mel-Spectrograms Generation Pipeline.

3.1.2 ASVspoof LA Datasets

The ASVspoof 2019 [32] and 2021 [38] datasets are designed to benchmark ASV systems against spoofing attacks. ASVspoof 2019 focuses on LA attacks using synthetic speech generated by 17 TTS and VC systems, with genuine utterances from 107 English-speaking speakers, recorded at 16 kHz in FLAC format and divided into training, development, and evaluation subsets. ASVspoof [38] significantly expands on this by including both logical and physical access (PA) scenarios, introducing more sophisticated spoofing attacks such as those based on modern neural vocoders and replay attacks captured under varied real-world conditions. It encompasses a diverse range of acoustic environments, including various playback/recording devices, background noise, and room acoustics. The 2021 [38] challenge supports two evaluation tracks, i.e, LA for synthetic speech detection and PA for replay attack detection, which offer a more rigorous and realistic assessment of ASV system robustness. Detailed statistics and dataset configurations for both editions are summarized in Table 1.

3.1.3 Urdu-DF Corpus

The Urdu deepfake corpus [27] comprises genuine, successfully trained voice samples and spoofing disturbances generated by two TTS systems: Tacotron and VITS. This dataset consists of clean audio from 17 speakers (7 female and 10 male) recorded in a professional studio. The participants recorded 708 sentences from the Phonetically Rich Urdu Speech (PRUS) corpus and 495 sentences from the news corpus [27]. Deepfake samples were generated exclusively for the news corpus, with 495 samples synthesized using each of the Tacotron and VITS models. The public corpus contains almost 200 bona fide and 200 spoofed samples per speaker, totaling 6,794 audio samples. A complete summary of dataset statistics is provided in Table 1.

3.2 Preprocessing: Spectrogram Generation

Figure 1 illustrates the Mel spectrogram generation pipeline for 2D audio signal representation from raw audio files (FLAC format, 16 kHz sampling rate). Each audio file is loaded and converted from stereo to mono if needed. To ensure consistency across samples and manage memory efficiently, we limit all audio to 10 seconds, applying truncation or zero-padding as appropriate. Mel-spectrograms are computed with 128 mel bins using a 1024-point FFT and a hop length of 512. The power spectrograms are converted to the decibel scale via amplitude-to-dB transformation. The resulting spectrogram matrix is visualized as a 256×256 image using the Inferno colormap, without axis labels, and saved as PNG files. We organize these outputs by dataset split (train, dev, eval) and class label (bona fide,

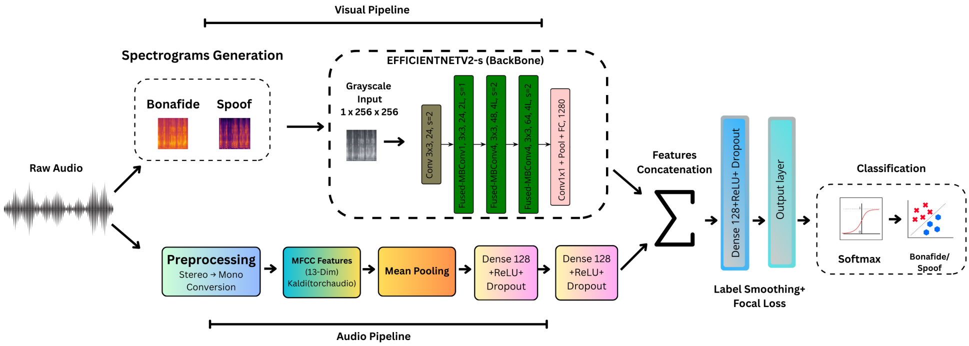

Figure 2: Overview of the proposed MFCC-EffNet architecture.

spoof). The entire process is optimized for batch processing using PyTorch and Torchaudio, allowing for efficient generation of large-scale spectrogram datasets.

3.3 Feature Extraction: MFCC Computation

Our framework uses a dedicated MFCC extraction pipeline to generate compact, perceptually relevant features for spoof detection. MFCCs are effective in speech processing due to their ability to mimic the human auditory system [2]. As shown in Figure 2, 16 kHz audio is first standardized and converted to mono and padded or truncated to 10 seconds. It is then segmented using a 25 ms Hamming window with a 10 ms hop size. Each frame yields a 13-dimensional MFCC vector computed via short-time Fourier transform (STFT), followed by Mel filterbank analysis, logarithmic compression, and a Discrete Cosine Transform (DCT), as defined by:

$$\text{MFCC}_n = \sum_{k=1}^{K} \log(E_k) \cdot \cos\left[\frac{\pi n(k-0.5)}{K}\right], \quad n = 1, 2, \ldots, N \tag{1}$$

where E_k is the output of the k^{th} Mel filter, K is the number of filters, and N is the number of coefficients (typically 13). We apply mean pooling across time to produce a fixed-size embedding, which is processed by two fully connected layers (128 and 64 units, both with ReLU and dropout) to enhance discriminability. The entire pipeline is implemented using a Kaldi-compatible extractor with GPU-accelerated Torchaudio for scalable batch processing.

3.4 Feature Extraction: Visual Computation

For the visual feature extraction module, EfficientNetV2-S [30] is used as a backbone model, as illustrated in Figure 2. EfficientNetV2-S [30] is a highly efficient CNN architecture that balances accuracy and computational cost via training-aware neural architecture search through compound scaling of depth, width, and resolution. In our modified architecture, the first convolutional layer of EfficientNetV2-S is adapted to accept single-channel (grayscale) spectrogram images (of shape 1×296×296) instead of the standard three-channel RGB input. This enables the model to directly process the spectrograms without requiring color transformations. We initialize the backbone using pre-trained ImageNet weights and fine-tune it on our task-specific dataset to improve convergence. To reduce computational load and avoid redundancy in high-level visual features, we truncate the EfficientNetV2-S backbone to retain only the first five stages, up to the 64-channel Fused-MBConv4 block. These layers include an initial 3×3 convolution (stride 2, 24 channels), two Fused-MBConv1 blocks (24 channels), four Fused-MBConv4 blocks (stride 2, 48 channels), and another four Fused-MBConv4 blocks (64 channels, stride 2). This truncation limits the total parameter count to under 250K, offering a favorable trade-off between model efficiency and performance in our spectrogram-based classification task. As shown in Figure 2, the extracted visual features from this truncated backbone are then concatenated with audio features before classification.

3.5 Classification

After extracting features from both the visual and audio branches, the resulting embeddings are concatenated to form a unified feature vector. This joint representation captures complementary information from both the spectrogram images and the MFCCs. The combined vector is passed through an additional dense layer with 128 units, ReLU activation, and dropout, followed by the final output layer that produces logits for binary classification (bona fide vs. spoofed speech). A softmax activation function is applied to obtain class probabilities, as represented in Figure 2 in the classification block. To address the class imbalance in the dataset and improve the model's focus on harder-to-classify examples, a label-smoothing focal loss is used. This loss function helps prevent over-confidence in the model's predictions while placing greater emphasis on samples that lie near the decision boundary, ultimately enhancing the system's robustness in detecting spoofing attacks.

4 EXPERIMENT

4.1 Performance Metrics

We evaluated the system using standard metrics: precision, recall, accuracy, confusion matrix, and equal error rate (EER). Precision is the ratio of correctly predicted positives to all predicted positives:

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{2}$$

Recall measures the proportion of actual positives correctly identified:

$$\text{Recall} = \frac{TP}{TP+FN}, \tag{3}$$

Accuracy reflects the overall correctness of predictions:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{4}$$

The confusion matrix summarizes classification outcomes:

$$\text{Confusion Matrix} = \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}, \tag{5}$$

Dataset	EER	Accuracy	Bona fide (Prec. / Rec. / F1)			Spoof (Prec. / Rec. / F1)			TP$_{Bona}$	FN$_{Bona}$	TP$_{Spoof}$	FP$_{Spoof}$
Proposed (Urdu)	0.53%	99.79%	0.990	0.984	0.987	0.998	0.999	0.998	5345	86	59,450	50
Proposed (Hindi)	0.81%	99.45%	0.910	0.973	0.941	0.999	0.994	0.996	3243	85	84,655	482
Urdu-DF	0.41%	99.17%	0.988	0.991	0.990	0.988	0.985	0.986	595	5	593	7
ASVspoof 2019 (LA)	2.04%	96.14%	0.892	0.915	0.903	0.967	0.954	0.960	6,729	626	60,977	2,905
ASVspoof 2021 (LA)	5.14%	89.66%	0.812	0.823	0.817	0.897	0.871	0.884	12,194	2,622	116,236	17,124

Table 2: MFCC-EffNet performance across multiple datasets.

Finally, the equal error rate (EER) denotes the operating point where the false acceptance rate (FAR) equals the false rejection rate (FRR):

$$\text{EER} = \text{FAR}(\tau) = \text{FRR}(\tau), \tag{6}$$

with τ representing the decision threshold. Together, these metrics capture system accuracy, robustness, and trade-offs in spoof detection.

4.2 Experimental Setup

All experiments were conducted using PyTorch on a workstation equipped with two NVIDIA RTX 5000 GPUs, each providing 16GB of VRAM. The proposed MFCC-EffNet architecture integrates dual input branches: a visual stream based on EfficientNetV2-S and an audio stream based on MFCC features. The proposed dataset served as the primary benchmark; the training set was used for model fitting, the development set for validation and model weight updates, and the evaluation set for final performance assessment with dataset partitioning strictly following the protocol detailed in Table 1. We employed the AdamW optimizer with a learning rate of 10^{-3} and weight decay of 10^{-4}. A learning rate scheduler (ReduceLROnPlateau) was used to adaptively reduce the learning rate upon plateauing validation loss, with a factor of 0.5 and patience of 3 epochs. To address class imbalance, training used a combination of label-smoothing focal loss ($\alpha = 0.9$, $\gamma = 2$, smoothing factor 0.05) and weighted cross-entropy loss, where class weights were computed inversely proportional to class frequencies. Balanced sampling further ensured equitable class representation during training. Mixed precision training with automatic gradient scaling was enabled for computational efficiency. The model was trained for 25 epochs with a batch size of 128, and the checkpoint with the lowest equal error rate (EER) on the development set was selected for final evaluation.

4.3 Results & Discussions

Table 2 presents a systematic evaluation of the MFCC-EffNet model across five test sets: the Urdu and Hindi sets of the proposed multilingual corpora, two ASVspoof benchmarks, and a low-resource deepfake scenario. On the proposed Urdu set, the model maintains high efficacy (EER=0.53%) indicating that spectrogram and MFCC features learned effectively to a linguistic context without overfitting. The model shows a modest decline (EER=0.81%) on the proposed Hindi set likely from overlapping spectral characteristics. On the Urdu deepfake corpus containing TTS attacks (VITS TTS & Tacotron) [27], MFCC-EffNet achieves 0.41% EER, highlighting robust transferability to low-resource deepfake scenarios. In contrast, ASVspoof 2019 LA [32] yields 2.04% EER with 2,905 false alarms, and ASVspoof

Model	Architecture	ASVspoof 2019 (LA)	ASVspoof 2021 (LA)	Urdu-DF [27]
MFCC-EffNet (This work)	2D CNN / MFCC + spectrogram	**2.04**	**5.14**	**0.41**
Hybrid CNN-LSTM [25]	2D CNN + LSTM / MFCC + CQCC	2.20	—	—
MFCC + CNN (Baseline)[58]	MFCCs as 2D input to CNN	6.73	12.60	—
RawNet2 [14]	1D CNN / Raw waveform	4.61	8.36	0.51
Spectrogram ResNet41 [4]	2D CNN / Mel + Gammatone spectrogram	1.70	0.50	—
AASIST-L [35]	1D CNN / Temporal Convolutions	—	—	0.5

Table 3: Comparison with existing approaches in terms of EER (%).

2021 LA (14,816 / 133,360) shows 5.14% EER with over 17,000 misclassified spoofs. This emphasizes the need for domain-robust training, e.g., adversarial regularization, dynamic augmentation, and self-supervised pretraining.

Table 3 compares MFCC-EffNet performance with recent spoof detectors. MFCC-EffNet substantially outperforms MFCC-only CNN baselines (6.73%, 12.60%) and the Hybrid CNN–LSTM [25] (2.20% on 2019 LA). Against raw-waveform systems like A-RawNet2 (4.61%, 8.36%), MFCC-EffNet shows a clear advantage. Spectrogram-ResNet41 [4] achieves strong scores (1.70% 2019 LA, 0.50% 2021 DF) but its 10M-parameter backbone risks overfitting and requires careful regularization. In contrast, MFCC-EffNet's convolution architecture produces smoother training gradients, robustness to unseen attacks, and inference under 5ms/sample. On the Urdu-DF corpus, AASIST-L [35] achieves 0.50% EER, while MFCC-EffNet attains 0.41% without task-specific tuning, showing superior generalization to diverse deepfake methods. MFCC-EffNet outperforms hybrid CNN-LSTM [28] (2.2% EER) and their model required 10× higher computational cost, underscoring MFCC-EffNet's efficiency–performance advantage. Overall, the proposed model takes advantage of both 1D and 2D audio representations by using MFCCs to encode phonetic structures, while spectrogram embeddings capture residual high-frequency noise and phase distortions.

5 Conclusion

This work introduced a multilingual spoofed speech dataset and MFCC-EffNet, a compact spoof detector that fuses MFCC descriptors with truncated EfficientNetV2 spectrogram embeddings through cross-attention. The framework achieves state-of-the-art performance across both controlled multilingual datasets and large-scale benchmarks, demonstrating strong generalization to diverse spoofing techniques. Notably, MFCC-EffNet delivers robust detection on Urdu and Hindi speech under low-resource and code-switching conditions, while remaining competitive on ASVspoof benchmarks. These results highlight the importance of domain-adapted feature fusion in addressing the distinct challenges of South Asian speech, including rich phonetic inventories and variable prosody. Looking ahead, future directions include self-supervised multilingual pretraining, multi-task modeling of code-switching dynamics, and adversarial feature regularization to enhance resilience against unconstrained and evolving attacks.

6 Acknowledgment

This work is supported by the Higher Education Commission, Pakistan [Grant number: 87309] and the Engineering & Physical Sciences Research Council, UK [Grant number: EP/Y002288/1].

References

[1] ASVspoof challenge. https://www.asvspoof.org/. Accessed Online: 22-01-2025.

[2] Zrar Kh. Abdul and Abdulbasit K. Al-Talabani. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, 10:122136–122158, 2022.

[3] Marinus Baas, Max de Boer, Marco Wiering, and David M.J. Tax. Voice conversion with just nearest neighbors. *arXiv preprint arXiv:2305.18975*, 2023. URL https://arxiv.org/abs/2305.18975.

[4] Nidhi Chakravarty and Mohit Dua. Publicly available datasets analysis and spectrogram-resnet41 based improved features extraction for audio spoof attack detection. *International Journal of Systems Assurance Engineering and Management*, 15 (12):5611–5636, 2024. doi: 10.1007/s13198-024-02550-1.

[5] Ke Chen, Xiaorui Du, Bowen Zhu, Zhiyao Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE, 2022.

[6] Hyeongjun Choi et al. BC-ResMax: Residual network with max feature map activation for spoofing detection. *arXiv preprint*, 2023.

[7] Sung Hyun Choi, Jong Ho Lee, and Nam Soo Kim. Diff-HierVC: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation. In *Proc. Interspeech*, 2023. URL https://arxiv.org/abs/2311.04693.

[8] Mohit Dua, Swati Meena, Neelam, Amisha, and Nidhi Chakravarty. Audio deepfake detection using data augmented graph frequency cepstral coefficients. In *2023 International Conference on Sustainable Computing and Smart Systems (ICSCAN)*, pages 1–6, 2023. doi: 10.1109/ICSCAN58655.2023.10395679.

[9] PA Tamayo Flórez, Rubén Manrique, and B Pereira Nunes. HABLA: A dataset of Latin American Spanish accents for voice anti-spoofing. In *Proc. INTERSPEECH*, volume 2023, pages 1963–1967, 2023.

[10] Yuchen Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

[11] Yuchen Gong, Chien-I Lai, Yu-An Chung, and James Glass. SSAST: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10699–10709, 2022.

[12] Lazaro Gonzalez-Soler, Marta Gomez-Barrero, Madhu Kamble, Massimiliano Todisco, and Christoph Busch. Dual-stream temporal convolutional neural network for voice presentation attack detection. In *2022 International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2022. doi: 10.1109/IWBF55382.2022.9794518.

[13] Hafsa Habib, Huma Tauseef, Muhammad Abuzar Fahiem, Saima Farhan, and Ghousia Usman. Speakernet for cross-lingual text-independent speaker verification. *Archives of Acoustics*, 47(2):191–199, 2022. doi: 10.24425/aoa.2022.141641.

[14] Hye-jin Heo, Jee-weon Jung, Hee-Soo Shim, and Ha-Jin Yu. RawNet2: Learning raw audio features for end-to-end spoofing detection with graph attention networks. In *Interspeech*, pages 1418–1422, 2023.

[15] Lian Huang and Chi-Man Pun. Self-attention and hybrid features for replay and deep-fake audio detection. *arXiv preprint arXiv:2401.05614*, 2024.

[16] Jee-weon Jung, Hee-Soo Heo, Ju-ho Kim, Hye-jin Shim, and Ha-Jin Yu. RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification, 2019. URL https://arxiv.org/abs/1904.08104.

[17] Awais Khan and Khalid Mahmood Malik. Securing voice biometrics: One-shot learning approach for audio deepfake detection, 2023. URL https://arxiv.org/abs/2310.03856.

[18] Il-Youp Kwak et al. Low-quality fake audio detection through frequency feature masking: DDWS-Conv and BC-ResMax evaluation. *arXiv preprint*, 2024.

[19] Sang-Hoon Lee, Ha-Yeong Choi, Seung-Bin Kim, and Seong-Whan Lee. Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis. *arXiv preprint arXiv:2311.12454*, 2023.

[20] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. Audio anti-spoofing detection: A survey. *arXiv preprint arXiv:2404.13914*, 2024.

[21] Tu Li and Junzhe Xiao. FreeVC: Towards high-quality text-free one-shot voice conversion, 2022. URL https://arxiv.org/abs/2210.15418.

[22] Songting Liu. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*, 2024.

[23] Haoxin Ma, Jiangyan Yi, Chenglong Wang, Xinrui Yan, Jianhua Tao, Tao Wang, Shiming Wang, and Ruibo Fu. CFAD: A Chinese dataset for fake audio detection. *Speech Communication*, 164:103122, 2024.

[24] Juan M. Martín-Doñas and Alfonso Álvarez. Enhancing voice spoofing detection models with wav2vec 2.0. *arXiv preprint*, 2023.

[25] Neelima Medikonda and I. Santi Prabha. Hybrid feature optimization for voice spoof detection using CNN-LSTM. *Traitement du Signal*, 41(2):717–727, 2024. doi: 10.18280/ts.410214.

[26] Nicolas M Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. Mlaad: The multi-language audio anti-spoofing dataset. *arXiv preprint arXiv:2401.09512*, 2024.

[27] Sheza Munir, Wassay Sajjad, Mukeet Raza, Emaan Mujahid Abbas, Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza. Deepfake defense: Constructing and evaluating a specialized urdu deepfake audio dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 2024.

[28] Pranita Palsapure, Rajeswari Rajeswari, Sandeep Kempegowda, and Kumbhar Ravikumar. Discriminative deep learning based hybrid spectro-temporal features for synthetic voice spoofing detection. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 14(1):130–141, 2025. doi: 10.11591/ijai.v14.i1.pp130-141.

[29] Muhammad Saad Saeed, Shah Nawaz, Marta Moscati, Rohan Kumar Das, Muhammad Salman Tahir, Muhammad Zaigham Zaheer, Muhammad Irzam Liaqat, Muhammad Haris Khan, Karthik Nandakumar, Muhammad Haroon Yousaf, and Markus Schedl. A synopsis of fame 2024 challenge: Associating faces with voices in multilingual environments. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 11333–11334, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3688978. URL https://doi.org/10.1145/3664647.3688978.

[30] Mingxing Tan and Quoc Le. EfficientNetV2: Smaller models and faster training, 2021.

[31] Massimiliano Todisco, Humberto Delgado, and Nicholas Evans. A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients. In *Speaker Odyssey Workshop*, Bilbao, Spain, 2016.

[32] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. ASVspoof 2019: Future horizons in spoofed and fake speech detection. In *Interspeech*, pages 1008–1012, 2019.

[33] Hoan My Tran, David Guennec, Philippe Martin, Aghilas Sini, Damien Lolive, Arnaud Delhay, and Pierre-François Marteau. Spoofed speech detection with a focus on speaker embedding. In *Interspeech 2024*, pages 2080–2084, 2024. doi: 10.21437/Interspeech.2024-481.

[34] Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, et al. ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. *arXiv preprint arXiv:2408.08739*, 2024.

[35] Jee weon Jung, Hee-Soo Heo, Hemlata Tak, Hye jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks, 2021. URL https://arxiv.org/abs/2110.01200.

[36] Jee weon Jung, You Jin Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. Pushing the limits of raw waveform speaker recognition, 2022. URL https://arxiv.org/abs/2203.08488.

[37] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, and Junichi Yamagishi. ASVspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *Training*, 10(15):3750, 2014.

[38] Junichi Yamagishi, Massimiliano Todisco, Md Sahidullah, Héctor Delgado, Xin Wang, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, Ville Vestman, and Andreas Nautsch. ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection, 2021. URL https://arxiv.org/abs/2109.00537.

[39] Zhenyu Zhang, Yewei Gu, Xiaowei Yi, and Xianfeng Zhao. FMFCC-a: A challenging Mandarin dataset for synthetic speech detection. In *International Workshop on Digital Watermarking*, pages 117–131. Springer, 2021.