Momenta Internship Assignment

Audio Deepfake Detection

Name: Mohammed Sahil Rizvi

Email: msahilrizvi@gmail.com

Phone no.: +91-9601341856

Identified Approaches

1. DNN based spoofing detection [1]

- A Deep Neural Network (DNN) was used as a feature extractor as part of the ASVspoof challenge 2015.
- Uses context-augmented input frames.
- A final representation of the audio was achieved, the spoofing vector (s-vector).
- Mahalanobis distance and normalization methods like PLDA, test normalization (TNorm) and probabilistic normalization (PNorm) are investigated to get the best system performance.
- Applied on a dataset with 6 labels: : human, S1, S2, S3, S4, S5 (5 algorithms given in training set)
- Achieved the following results:

Table 4: Final result (EER(%)) on evaluation data

# Target	Norm	Known	Unknown	All
6	TNorm	0.058	4.998	2.528
6	PNorm	0.046	4.516	2.281
6	PLDA	8.650	20.54	14.59

- It is suited for Momenta's use case for the following reasons:
  - Robus to unseen spoofing methods
  - Frame-level processing and averaging make it suitable for streaming or near-real-time inference.
- Potential challenges/limitations
  - Trained on clean, segmented utterances — adaptation may be required to handle noisy, spontaneous, or multi-speaker conversations.
  - Lack of conversational dynamics in the training data; might need retraining or fine-tuning on dialogue-style speech for real-world use cases.

2. Res2Net architecture [2]

- Uses Res2Net for multi-scale feature representation within a single residual block by splitting input channels into several groups and applying hierarchical residual-like connections.
- SE (Squeeze-and-Excitation) Block Integration channels interdependencies to dynamically recalibrate features, helping focus on spoofing-relevant information
- Fusion Strategy: Uses a late fusion of multiple feature extractors (Spec, LFCC, CQT) to exploit complementary information across different audio representations.

- Performance comparison with state-of-the-art systems

System	Physical Access		Logical Access	
	EER (%)	t-DCF	EER (%)	t-DCF
Spec+ResNet+CE [13]	3.81	0.0994	9.68	0.2741
MFCC+ResNet+CE [13]	–	–	9.33	0.2042
CQCC+ResNet+CE [13]	4.43	0.1070	7.69	0.2166
Spec+ResNet+CE [15]	1.29	0.036	11.75	0.216
Joint-gram+ResNet+CE [14]	1.23	0.0305	–	–
GD-gram+ResNet+CE [14]	1.08	0.0282	–	–
LFCC+LCNN+A-softmax [17]	4.60	0.1053	5.06	0.1000
FFT+LCNN+A-softmax [17]	–	–	4.53	0.1028
CQT+LCNN+A-softmax [17]	1.23	0.0295	–	–
FG-CQT+LCNN+CE [18]	–	–	4.07	0.102
Spec+LCGRNN+GKDE-Softmax [16]	1.06	0.0222	3.77	0.0842
Spec+LCGRNN+GKDE-Triplet [16]	0.92	0.0198	3.03	0.0776
MGD+ResNeWt+CE [11]	2.15	0.0465	–	–
CQTMGD+ResNeWt+CE [11]	0.94	0.0250	–	–
Fbanks&CQT+ResNeWt+CE [11]	0.52	0.0134	–	–
Ours: CQT+SE-Res2Net50+CE	**0.459**	**0.0116**	**2.502**	**0.0743**

- Suitable for use-case:
  - Handles unseen spoofing attacks better than deeper or traditional architectures like ResNet34/50.
  - Compatible with multiple audio features and benefits from feature fusion.
  - Has smaller model size (e.g., Res2Net50 has fewer parameters than ResNet50) while improving performance.
- Limitations:
  - Combining multiple models adds computational cost at inference time.
  - Fusion and SE-Res2Net50 may need optimization for low-latency or edge deployment.

3. Graph-based model [3]

- Divides T-F representations into overlapping patches for richer feature extraction.
- Constructs a graph over patches, connecting nodes that share time or frequency characteristics.
- Computes weighted edges using patch similarity via a learnable projection network
- Utilizes a graph convolutional network (GCN) to propagate contextual information.
- Performance results

System	min t-DCF	EER(%)
NP	0.0314	1.24
FC	0.0281	1.05
ST	0.0539	1.82
SF	0.0304	1.20
UW	0.0600	2.07
UW-FC	0.0593	2.00
AP	0.0377	1.44
MP	0.0326	1.21
NC	0.0320	1.18
Proposed	**0.0166**	**0.58**

- Suitability for use-case:
  - The model's patch-level processing and graph structure capture fine-grained temporal and spectral inconsistencies, which are typical artifacts in synthetic or generated audio.
  - The architecture uses a relatively shallow GCN with low-dimensional embeddings, which suggests potential for near real-time detection, especially if optimized or pruned for deployment.
- Limitations:

- For long utterances or high-resolution features, building the graph and computing adjacency weights could be computationally expensive.
- Edge definition depends on heuristics (same subband/time segment), which might limit adaptability to more complex or noisy signals.

Resources:

[1] Chen, Nanxin & Qian, Yanmin & Dinkel, Heinrich & Chen, Bo & Yu, Kai. (2015). Robust deep feature for spoofing detection — the SJTU system for ASVspoof 2015 challenge. 2097-2101. 10.21437/Interspeech.2015-474.

[2] X. Li et al., "Replay and Synthetic Speech Detection with Res2Net Architecture," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 6354-6358, doi: 10.1109/ICASSP39728.2021.9413828.

[3] Chen, Feng & Deng, Shi-wen & Zheng, Tieran & He, Yongjun & Han, Jiqing. (2023). Graph-Based Spectro-Temporal Dependency Modeling for Anti-Spoofing. 1-5. 10.1109/ICASSP49357.2023.10096741.