Uncovering the Deceptions: An Analysis on Audio Spoofing Detection and Future Prospects Rishabh Ranjan , Mayank Vatsa , Richa Singh - Indian Institute of Technology, Jodhpur, India

I picked this paper as I am familiar with the institution and the paper is extremely precise. It also covers some of the most common Audio attacks and efficiency metrics of different models while addressing these attacks. I am using this paper as a foundation.

Detection methods

- Initially Machine Learning approaches - Gaussian mixture model, Hidden Markov model, Support Vector Machines were commonly utilized before 2015
- Deep Learning approaches Convolutional Neural Networks and Recurrent Neural Networks have been popularized recently.

Takeaways

- This paper has mentioned error metrics for multiple approaches on ASVspoof2019LA.
- ASSIST and ASSIST-L have shown high performance against TTS and VC attacks while being cross-linguistics
- Face+SE-Res2Net and Spec+LCGRNN have shown robust performance against Replay attacks

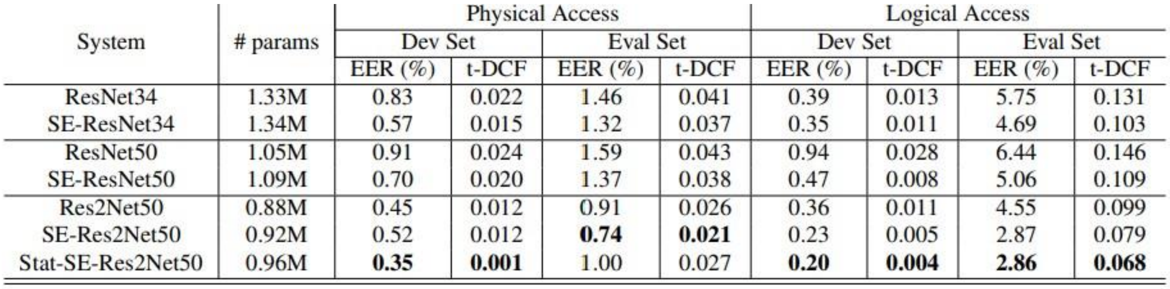
Challenges

- Privacy policies must be maintained (General Data Protection Regulation) while developing these datasets as they are going to require large amounts of data entities to train accuracy for these detection models.
- For a diverse country like India, the datasets need to be diversely tuned towards different dialects, accents and languages.
- The quality of data must be maintained.

Replay and Synthetic Speech Detection with Res2Net Architecture Xu Li; Na Li; Chao Weng; Xunying Liu; Dan Su; Dong Yu

ResNet stands for residual networks. I have worked with Res2Net in the computer vision domain, and I understand how they work. These are combined with Neural Networks in the audio domain to detect specific patterns in spectrograms. Res2net50 can also be Squeeze and excite block and Constant Q transform (CQT) for promising results.

Takeaways

- ResNet involves studying audio spectrums as it solves the vanishing gradient problems. Constant Q-transform provides high resolution at low freq.
- Res2net50 has shown great performance against Replay and Synthetic Speech attacks
- 

System	# params	Physical Access				Logical Access			
		Dev Set		Eval Set		Dev Set		Eval Set	
		EER (%)	t-DCF	EER (%)	t-DCF	EER (%)	t-DCF	EER (%)	t-DCF
ResNet34	1.33M	0.83	0.022	1.46	0.041	0.39	0.013	5.75	0.131
SE-ResNet34	1.34M	0.57	0.015	1.32	0.037	0.35	0.011	4.69	0.103
ResNet50	1.05M	0.91	0.024	1.59	0.043	0.94	0.028	6.44	0.146
SE-ResNet50	1.09M	0.70	0.020	1.37	0.038	0.47	0.008	5.06	0.109
Res2Net50	0.88M	0.45	0.012	0.91	0.026	0.36	0.011	4.55	0.099
SE-Res2Net50	0.92M	0.52	0.012	**0.74**	**0.021**	0.23	0.005	2.87	0.079
Stat-SE-Res2Net50	0.96M	**0.35**	**0.001**	1.00	0.027	**0.20**	**0.004**	**2.86**	**0.068**

Challenges

- Hardware requirements are relatively heavier and may require separate optimization.
- Performance against Adversarial attacks is not up to par with the rest of the metrics.
- Performance is limited if training datasets are not diverse.

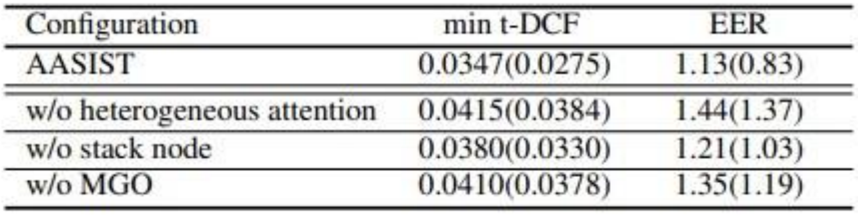
AASIST: AUDIO ANTI-SPOOFING USING INTEGRATED SPECTRO-TEMPORAL GRAPH ATTENTION NETWORKS

Jee-weon Jung1 , Hee-Soo Heo1 , Hemlata Tak2 , Hye-jin Shim3 , Joon Son Chung1 , Bong-Jin Lee1 , Ha-Jin Yu3 , Nicholas Evans2

AASIST is based on the premise that bona-fide and spoofed utterances can be differentiated by judging the Spectro-Temporal patterns. This system utilized the RawNet2 based encoder.

Takeaways

- With AASIST we can judge and differentiate subtle artifacts introduced to the data through AI.
- Through utilization of Graphical attention theory, we can employ both speaker verification and spoofing detection.
- For limited environments, an identical lightweight version named AASIST-L can be employed.



Configuration	min t-DCF	EER
AASIST	0.0347(0.0275)	1.13(0.83)
w/o heterogeneous attention	0.0415(0.0384)	1.44(1.37)
w/o stack node	0.0380(0.0330)	1.21(1.03)
w/o MGO	0.0410(0.0378)	1.35(1.19)

Challenges

- Heavy hardware requirements during training.
- Susceptible to overtraining and becoming biased towards the training dataset.

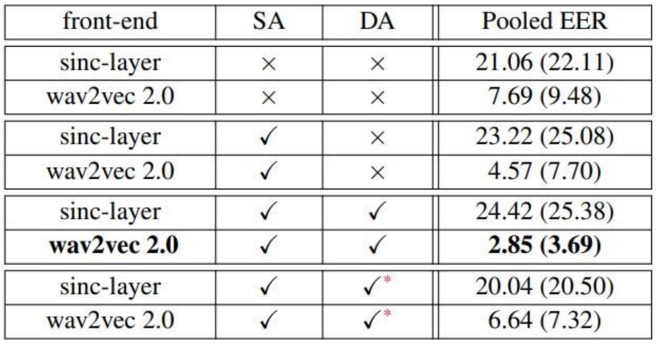
Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation

Hemlata Tak , Massimiliano Todisco , Xin Wang2, Jee-weon Jung, Junichi Yamagishi and Nicholas Evans

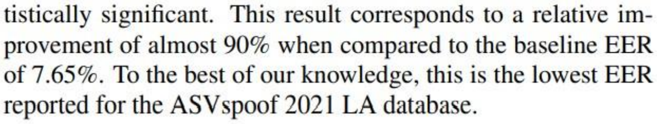
This approach is based on self-supervised learning through wav2vec 2.0. Trained only on bona-fide data, this approach has recorded some of the lowest error rates in the ASVspoof2021 Logical access and Deepfake databases

Takeaways

- The model is strictly trained on genuine data.
- Data Augmentation is utilized to reduce overfitting to get even better results.
- Wav2vec is integrated by directly feeding raw waveforms to the system.



front-end	SA	DA	Pooled EER
sinc-layer	×	×	21.06 (22.11)
wav2vec 2.0	×	×	7.69 (9.48)
sinc-layer	✓	×	23.22 (25.08)
wav2vec 2.0	✓	×	4.57 (7.70)
sinc-layer	✓	✓	24.42 (25.38)
wav2vec 2.0	✓	✓	**2.85 (3.69)**
sinc-layer	✓	✓*	20.04 (20.50)
wav2vec 2.0	✓	✓*	6.64 (7.32)



tistically significant. This result corresponds to a relative improvement of almost 90% when compared to the baseline EER of 7.65%. To the best of our knowledge, this is the lowest EER reported for the ASVspoof 2021 LA database.

Challenges

- Since the training data is strictly genuine, the model may struggle against anomalies.
- The models trained on certain datasets may perform better in confined scenarios compared to real life scenarios.