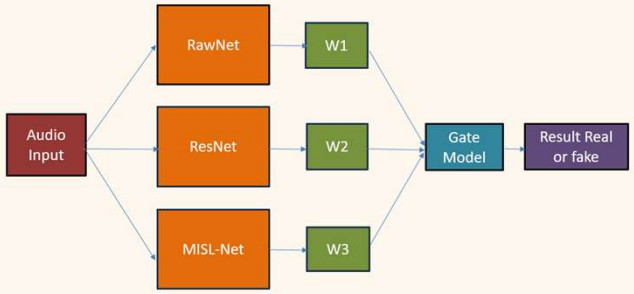
Synthetic Audio Detector (Team 29)

David Volchonok[1], Brenda Krishnawongso[1], Ryan Ma[1], and Dr. Stamm[1]

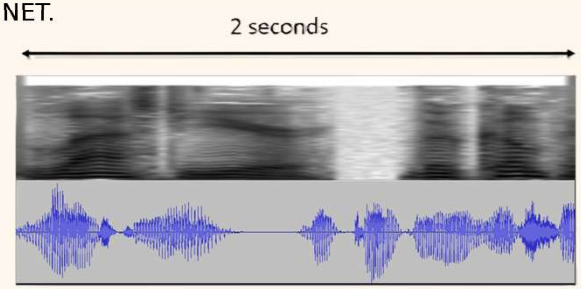
[1]Department of Electrical and Computer Engineering

Abstract

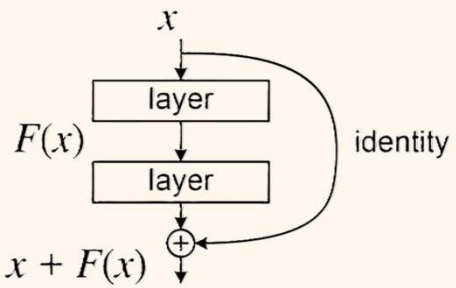
Digital advancements has introduced new security vulnerabilities, including voice phishing and other forms of audio-based fraud with significant harmful potential. Current authentication systems often fail to distinguish artificial speech from authentic human waveforms, leaving users vulnerable to increasingly convincing voice phishing. To address this problem, we present a Synthetic Audio Detector (SAD) which is easy for consumers to use.

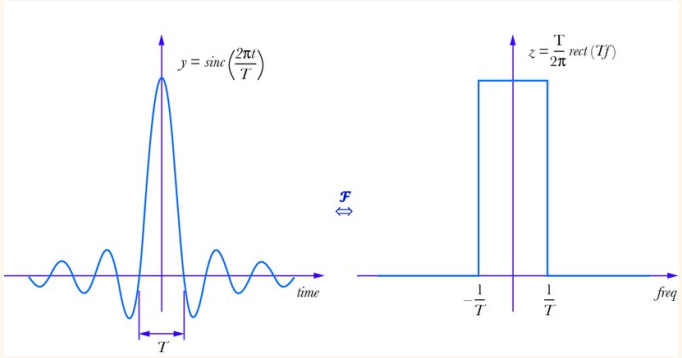


SAD is designed to distinguish between synthetic and authentic human speech. The proposed system employs an ensemble of **specialized expert models**, consisting of Residual-block based architectures as well as a simpler MISL-NET.



Models



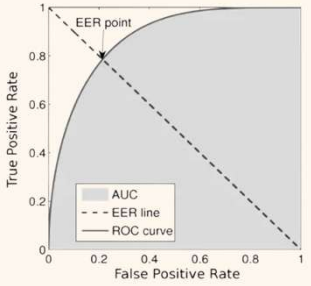


$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

Methodology

The Synthetic Audio Detector (SAD) was evaluated using the ASVspoof 2019, In the Wild, and FakeOrReal datasets which contain both authentic and synthetic speech samples. Data was converted into each the model's preferred format prior to testing. Model parameters were adjusted to achieve optimal performance through a series of parameter studies and architecture alterations.

Metrics



$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Results

Single Models

Model	AU-ROC	EER	F1 (thrs= 0.5)	Inference Time
ResNetMel	**0.9285**	**0.1188**	**0.8835**	17.4 ms
ResNetLin	0.9285	0.1273	0.8778	17.4 ms
RawNet	0.9075	0.1594	0.8464	107.3 ms
MISL-Net	0.9098	0.1634	0.8443	**1.5 ms**

Two Models

Model	AU-ROC	EER
RNM_RNL	**0.9688**	**0.0751**
RNM_RAW	0.9435	0.1051
RNM_MSL	0.9413	0.1171
RNL_RAW	0.9507	0.1051
RNL_MSL	0.9502	0.1021
MSL_RAW	0.9275	0.1394

Conclusion

The proposed SAD framework provides a scalable and consumer-accessible foundation for mitigating voice phishing and synthetic audio fraud. It is very easy to create a new model and install it into the system or to update an older one. As generative speech technologies continue to advance, future work will focus on improving generalization to unseen synthesis models, incorporating zero-shot strategies and optimizing deployment for real-time applications.

Further Developments

The UI could be updated to better explain how to use SAD. Additionally, the models are still imperfect and can be made faster through hosting on a more powerful machine.

misl multimedia & information security lab

Drexel UNIVERSITY