Audio Spoof Detection: A Comprehensive Analysis of Multi-Model Approaches in SafeSpeak-2024

Team: misis voice scoofing detectors
Ryzhichkin Kirill, Sergeev Daniil

***Abstract*—The escalating sophistication of audio spoofing techniques necessitates advanced detection methodologies. This study presents a comprehensive investigation of audio spoof detection models, leveraging state-of-the-art self-supervised learning (SSL) and transformer architectures. By analyzing multiple model architectures across the ASVspoof benchmark datasets, we demonstrate significant improvements in distinguishing between genuine and synthesized speech.**

I. Introduction

VOICE authentication systems face unprecedented challenges from increasingly sophisticated audio spoofing technologies. The proliferation of deep learning and generative models has created critical vulnerabilities in biometric security systems.

A. Research Objectives

Our primary research objectives include:

- Evaluating multiple deep learning architectures for audio spoof detection
- Analyzing computational efficiency and detection performance
- Developing robust methodologies for distinguishing synthetic from genuine speech

II. Theoretical Framework

A. Equal Error Rate Formulation

The Equal Error Rate (EER) represents a critical metric in binary classification for audio spoof detection. Mathematically, EER is defined as:

$$EER = \min_{t} \left\{ \frac{FAR(t) + FRR(t)}{2} \right\} \tag{1}$$

Where:

- t represents the classification threshold
- $FAR(t)$ is the False Acceptance Rate
- $FRR(t)$ is the False Rejection Rate

III. Methodology

A. Datasets

- ASVspoof 2019 (primary training dataset)
- ASVspoof 2021 (supplementary evaluation)

B. Model Architectures

We implemented and compared six distinct model architectures:

1) Wav2Vec 2.0
2) HuBERT
3) SSL Wav2Vec 2.0 with PSFAN Backend
4) Audio Spectral Transformer
5) Sound Event Detection Model (EfficientNet-B0)
6) WavLM Base

C. Data Augmentation Strategies

Advanced augmentation techniques were employed:

- Gaussian noise injection
- Signal-to-noise ratio modifications
- Dynamic gain variations
- Convolutive noise introduction
- Background noise injection

D. Loss Functions and Optimization

To enhance model performance and address class imbalance, we experimented with various loss functions, including Focal Loss, Cross-Entropy Loss, and BCEWithLogits Loss. For optimization, we used the AdamW optimizer, known for its efficiency in training transformer models. Learning rate scheduling was handled using linear and cosine schedulers, ensuring smooth convergence during training.

IV. Experimental Results

A. Performance Metrics

Table I presents the comprehensive performance evaluation.

TABLE I
Detailed Model Performance Comparison

Model	Public LB EER	Precision	Recall	F1-Score
Wav2Vec 2.0	0.46516	0.888	0.788	0.835
SSL Wav2Vec	0.02925	-	-	-
Audio Spectral Transformer	0.01384	0.999	0.999	0.999
HuBERT	8.11672	0.877	0.764	0.817
Pretrained Wav2Vec	0.77492	0.845	0.725	0.780
WavLM Base	1.87658	0.820	0.690	0.750

V. Discussion and Analysis

A. Performance Insights

From Table I, it is evident that the Audio Spectral Transformer and SSL-model significantly outperformed other models across all key metrics. The fact is that the AST and SSL models were trained on the full ASVSpoof dataset (train, eval, dev part), unlike the other models, which were trained only on the train part. The use of transformer architectures, known for their robust feature extraction capabilities, was a decisive factor in achieving near-perfect precision and recall. We also tried a greedy soup of models for SSL, but this did not give an improvement.

B. Impact of Data Augmentation

The incorporation of advanced data augmentation strategies played a pivotal role in model performance. For instance:

- **Gaussian Noise Injection:** Enhanced the model's ability to generalize to noisy environments, common in real-world scenarios.
- **Signal-to-Noise Ratio Modifications:** Allowed models to detect subtle patterns in low-quality audio.
- **Dynamic Gain Variations:** Provided robustness against varying audio amplitudes.
- **Background Noise Injection:** Simulated challenging acoustic conditions, ensuring the models performed well across diverse settings (not used in the final version).

Augmentation techniques contributed to a 3% improvement in final metric, particularly in transformer-based architectures.

C. Challenges and Limitations

Despite significant advancements, certain limitations were observed:

1) **Computational Overhead:** Transformer-based architectures, while powerful, require significant computational resources, potentially limiting deployment in resource-constrained environments (AST: 1 epoch training takes 4 hours, inference - 2 hours, SSL W2V: 1 epoch training takes 2.5 hours, inference - 1.8 hours; GPU - 2xT4).
2) **Generalization:** Although augmentation techniques and SOTA models improved performance, generalization to unseen spoofing methods remains a challenge.
3) **Model Interpretability:** The black-box nature of deep learning models makes it difficult to interpret decision-making processes, posing challenges for trustworthiness.

VI. Future Work

Based on the findings, several avenues for future research are proposed:

- **Ensemble Learning:** Combining the strengths of multiple models (e.g., transformers and SSL models) to enhance robustness.
- **Lightweight Architectures:** Developing computationally efficient variants of transformer-based models for deployment in real-time systems.
- **Adversarial Robustness:** Exploring defense mechanisms against adversarial attacks on voice authentication systems (maybe add TTA).

VII. Conclusion

This study has demonstrated the efficacy of state-of-the-art audio processing models in detecting spoofed audio, with transformer-based architectures, particularly the Audio Spectral Transformer, achieving near-perfect results. By leveraging advanced data augmentation techniques, the models exhibited strong generalization capabilities across diverse acoustic conditions. However, challenges such as computational overhead and limited generalization to novel spoofing methods highlight the need for ongoing innovation.

The findings underscore the critical importance of adopting advanced methodologies to safeguard voice authentication systems in an era of rapidly evolving audio spoofing technologies. Future research should focus on achieving a balance between performance and efficiency, ensuring the applicability of these systems in real-world scenarios.

Links

- Best Model Weights:
  - AST Spoofing Model (HuggingFace)
  - SSL Wav2Vec with PSFAN (HuggingFace)
- GitHub Repository: ASVSpoof Project Repository

Acknowledgments

The authors extend their gratitude to the organizers of SafeSpeak-2024 and the maintainers of the ASVspoof datasets for providing valuable resources that enabled this research.

References

[1] Kinnunen, T., Lee, K. A., Delgado, H., et al., "The ASVspoof 2019 challenge: TTS and VC spoofing attacks," *Proc. Interspeech*, 2019.
[2] Baevski, A., Zhou, H., Mohamed, A., Auli, M., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
[3] Tak, Hemlata, Todisco, Massimiliano, Wang, Xin, Jung, Jee-weon, Yamagishi, Junichi, and Evans, Nicholas, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," *The Speaker and Language Recognition Workshop*, 2022.