Audio Deepfake Detection

Dhannawat, Chakshu
B20AI006

Jangra, Vageesh
B20AI049

Shrivastava, Yash
B21CS079

1 Introduction

The proliferation of audio deepfake technology poses a significant threat to the authenticity and integrity of audio content in various domains, including security, media, and entertainment. With the advancement of machine learning techniques, particularly in the realm of speech processing, the creation of sophisticated audio deepfakes has become increasingly accessible, raising concerns about the potential misuse of such technology for malicious purposes.

In response to this growing threat, there is a pressing need for robust and effective methods to detect audio deepfakes and distinguish them from genuine audio recordings. By leveraging state-of-the-art models and feature extraction techniques, along with innovative loss functions, researchers aim to develop reliable audio deepfake detection systems capable of accurately identifying manipulated audio content.

This paper presents a comprehensive investigation into the detection of audio deepfakes, focusing on the evaluation of various models, feature extraction techniques, and loss functions. By examining the performance of these components on benchmark datasets such as ASVSpoof2021, the research aims to advance the current understanding of audio deepfake detection and contribute to the development of more resilient detection systems.
You can find our code in this GitHub Repository: Link

2 Problem Statement

The primary objective of this research is to address the challenge of detecting audio deepfakes amidst the increasing sophistication of manipulation techniques. Specifically, the study aims to:

- Evaluate the effectiveness of state-of-the-art models, including RawNet2, LCNN, for audio deepfake detection.
- Investigate the performance of feature extraction techniques such as MFCC and prosodic features, including F0, energy, and duration features, in combination with advanced models.
- Assess the impact of utilizing large margin cosine loss (LMCL) as an alternative loss function to enhance the robustness of feature embeddings in audio deepfake detection systems.

By addressing these objectives, the research seeks to contribute to the development of more reliable and resilient audio deepfake detection systems capable of mitigating the threat posed by manipulated audio content.

3 Literature Review

In recent years, significant progress has been made in the field of audio deepfake detection, with researchers proposing various models and techniques to address the evolving nature of audio manipulation. Some of the state-of-the-art models include RawNet2, LCNN, and AASIST, which have demonstrated promising results in detecting audio deepfakes.

Feature extraction plays a crucial role in audio deepfake detection, with techniques such as Mel-frequency cepstral coefficients (MFCC) and prosodic features being widely used for capturing distinctive characteristics of genuine and manipulated audio. The Yet Another Algorithm for Pitch Tracking method (YAPPT) has been utilized to extract F0 features, while HuBERT-based duration methods have been employed for duration features.

In addition to traditional feature extraction techniques, recent advancements have incorporated pretrained models such as Wav2vec2.0 XLS-R to enhance feature representations and improve detection performance. By combining prosodic features with XLS-R features, researchers have achieved notable improvements in detecting audio deepfakes.[3]

One of the key challenges in audio deepfake detection is the generalization ability of detection systems to unknown attacks. To address this challenge, researchers have explored alternative loss functions such as large margin cosine loss (LMCL). LMCL aims to optimize feature embeddings by maximizing interclass variance and minimizing intra-class variance, thereby improving the discriminative power of the model.[2]

Chen et al.[2] proposed LMCL as a cosine loss with a margin parameter to enhance the robustness of feature embeddings in deep neural networks. By reformulating the softmax loss as a cosine loss and introducing a margin in the cosine space, LMCL aims to improve the separability of features and enhance the performance of audio deepfake detection systems.

4 Methodology

4.1 Dataset

The experimentation was conducted using the ASVspoof2021 dataset, which contains audio recordings with both genuine and manipulated content. The dataset consists of a diverse set of speech samples collected from various sources and environments, providing a comprehensive evaluation platform for audio deepfake detection systems.

4.2 LMCL Loss

$$L_{lmc} = \frac{1}{N}\sum_{i} -log \frac{e^{s(cos(\theta_{y_i},i)-m)}}{e^{s(cos(\theta_{y_i},i)-m)} + \sum_{i \neq y_i} e^{scos(\theta_j,i)}} \quad (1)$$

subject to:

$$\begin{aligned} W &= \frac{W^*}{\|W^*\|}, \\ x &= \frac{x^*}{\|x^*\|}, \\ cos(\theta_j, i) &= W_j{}^T x_i. \end{aligned} \quad (2)$$

Figure 1: Mathematical Equation of LMCL loss

To enhance the robustness of the deepfake detection system, the large margin cosine loss (LMCL) was employed as an alternative loss function. LMCL aims to optimize feature embeddings by maximizing interclass variance and minimizing intra-class variance, thereby improving the discriminative power of the model. Figure 1 illustrates the concept of LMCL loss and its effectiveness in enhancing the separation between classes.

4.3 Experimentation

4.3.1 Performance of LCNN Model with Different Features on ASVspoof Dataset

Model	Features	EER (%)
LCNN	MFCC	32.91
LCNN	F0(A)	43.06
LCNN	Energy(B)	48.74
LCNN	Duration(C)	41.83
LCNN	XLS-R(Wav2Vec2)	30.32
LCNN	XLS-R+A+B+C	28.63

Table 1: Performance of LCNN model with different features on ASVspoof dataset

The experimentation on the ASVspoof2021 dataset using the LCNN model trained with binary cross-entropy (BCE) loss revealed varying performance across different feature sets. While MFCC and energy-based features exhibited relatively high equal-error rates (EER), the combination of XLS-R features with prosodic features (A+B+C) achieved the lowest EER of 28.63%, when LCNN model was used.

4.3.2 Comparisons of Loss Functions (BCE vs LMCL) - RawNet2 and LCNN Models on ASVspoof Dataset

Model	**Loss**	**EER (%)**
RawNet2	BCE	30.45
RawNet2	LMCL	29.68
LCNN	BCE	28.63
LCNN	LMCL	27.99

Table 2: Comparisons of loss functions (BCE vs LMCL) - RawNet2 and LCNN models on ASVspoof dataset

The investigation into alternative loss functions revealed that employing the large margin cosine loss (LMCL) resulted in a slight improvement in performance compared to the conventional BCE loss. The LMCL achieved an EER of 27.99, indicating its potential to enhance the robustness of the LCNN as well as RawNet2 models in detecting audio deepfakes. We have used the A+B+C+XLS-R feature pipeline as mentioned above, since they had the least EER, and as for RawNet2, it has its own preprocessing/feature extraction, so original audio files are used in it.
Based on the above experimentation, we can conclude that ir-respective of the features used, LMCL loss is able to increase the robustness of the audio deepfake models.

4.4 Critical Analysis

The critical analysis of the experimentation results revealed several key insights. Firstly, the combination of XLS-R features with prosodic features (A+B+C) achieved the lowest equal-error rate (EER), indicating improved performance in detecting audio deepfakes. Secondly, the adoption of the LMCL loss function led to a slight reduction in EER compared to the conventional cross-entropy (BCE) loss, highlighting its potential to enhance the robustness of deepfake detection systems. Lastly, further experimentation and optimization are warranted to explore additional feature combinations and loss functions for achieving even better performance.

5 Conclusion

The experimentation on audio deepfake detection using the ASVspoof2021 dataset and the LCNN model provided valuable insights into the effectiveness of different feature sets and loss functions. The results demonstrated that the combination of XLS-R features with prosodic features (A+B+C) significantly improved the model's performance, achieving the lowest EER. Additionally, the adoption of the LMCL loss function showed promising potential in further enhancing the robustness of deepfake detection systems. These findings underscore the importance of comprehensive experimentation and critical analysis in advancing the field of audio deepfake detection and mitigating the risks associated with manipulated audio content.

6 Individual Contributions

We ensured that each member contributed equally to the project, with collaborative efforts in coding and problem-solving throughout. The implementation and testing were not solely assigned to one person; instead, each of us worked together, resolving each other's errors along the way.

- **Chakshu:** Conducted literature review, implemented and tested RawNet2 model with various feature combinations.

- **Yash:** Conducted Literature Review, Implemented and tested LCNN model, Implemented script for feature extraction, also wrote the code for Loading ASVSpoof2021 dataset.

- **Vageesh:** Implemented and tested LMCL Loss, conducted testing of BCE vs LMCL losses. Also helped in testing the various feature combinations in terms of EER.

This collaborative approach ensured a well-rounded and comprehensive exploration of the project objectives.

7 References

1. Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao, "Audio Deepfake Detection: A Survey," *Journal of LaTeX Class Files*, vol. 14, no. 8, August 2023.

2. Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, and Elie Khoury, "Generalization Of Audio Deepfake Detection," in *The Speaker and Language Recognition Workshop (Odyssey 2020)*, 1-5 November 2020, Tokyo, Japan.

3. Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," arXiv:2111.13915 [cs.CL], version 3, submitted on 17 Nov 2021, last revised 16 Dec 2021.

4. ASVspoof dataset repository. Available online: `https://github.com/ASVSpoof/ASVspoof`