Spoofing and Countermeasures for Speaker Verification by Self-Supervised Learning

Subhajit Saha
Institute for Advancing Intelligence
TCG CREST, Kolkata, India-700091

November 2023

Abstract

Use of Automatic Speaker Verification (ASV) is a very common practice nowadays in the smart bio-metric systems. Over the last two decades Machine Learning (ML) as well as Deep Learning (DL) techniques has been adopted to ASV task using the acoustic features or raw waveform of speech data. But due to the recent advancement in spoofing techniques (i.e. by Deep fake) and more availability of original speaker data, distinguishing the fake audio with the bonafide speaker is more challenging than it was earlier. This research is a novel attempt to separate the bonafide speech data and spoofed speech data on a state of art dataset (ASVspoof 2019). Perhaps it is the first research that has enabled classical ML techniques in self supervised learning (SSL) approach for audio deep fake detection (ADFD). It makes the proposed model more computationally efficient, explainable and convenient to the AI community as well as speech researchers compared to the other state of art techniques. Moreover, a comprehensive study has also been conducted to explore the embedding quality of different layers from the used speech model (wav2vec 2.0). Finally this paper is concluded by mentioning possible research direction to modify the present technique as well as further challenges in ADFD or ASV.

1 Problem Statement

Automated Speaker Verification (ASV) is considered a fundamental requirement due to the widespread adoption of biometric authentication software in today's technology landscape. An ASV must be resilient to all kind of speech spoofing. Our goal is to build an AI model that detects three types of spoofing techniques.

1. **Text-to-Speech Synthesis (TTS):** - *Purpose:* Converts text to natural-sounding speech. - *Method:* Utilizes pre-designed voice models.
2. **Voice Conversion:** - *Purpose:* Modifies voice characteristics while retaining speech content. - *Method:* Analyzes and transforms source and target voices.
3. **Replay Spoofing:** - *Purpose:* Deceives voice recognition systems during authentication. - *Method:* Records and replays legitimate user's voice.

Deep learning based models are the most common approaches work on unstructured, high resolution and almost non separable manifold dataset. It has two fundamental limitations; Most of Deep learning models requires high volume of labeled data which may be very costly to generate or may not be available and training from scratch of high volume Deep Learning models are computationally expensive.

So we shall explore well used Machine Learning models to train and observe their performances. Not only it is computationally less expensive and generally requires less data to train also such models are easy to understand wheres still Deep Learning is kind of ”black box” model.

2 Dataset

We have used ASVspoof2019 dataset [3] which simulates all the three types of spoofing mentioned above. Precisely speaking they have provided two sub challenges. One for “Logical Access” scenario (LA; speech synthesis/conversion attacks) and another for “Physical Access” scenario (PA; replay attacks). We have experimented for LA only.

3 State of the art

As the evaluation of the proposed model we can use two metrics tandem decision cost function (t-DCF) and Equal error rate (EER) as dscribed by kinnunen et al. [1]. From the recorded leader-board of paper with code we have the following state of the art models.

Dataset	Rank	Model	EER	min t-DCF	Year
Voice Anti-spoofing on ASVspoof 2019 - LA	1	AASIST [4]	0.83	0.0275	2021
	2	LFCC&Face+SE-DenseNet+A-softmax [5]	2.73	0.0713	2023
Voice Anti-spoofing on ASVspoof 2019 - PA	1	logPowSpec+EABNet+CombLoss [2]	0.86	0.0239	2021

4 Methodology

1: Extract Embedding from wave2vec2.0 to get shape in $(N, 768)$ from each audio file of ASVspoof2019 dataset.
2: Make average to get shape 768 from each embedding.
3: Train ML algorithms after selecting hyperparameters based on the F1 score from validation dataset.
4: Evaluate each model on test dataset as shown on in Performance Metric table.

5 Performance Metric

[1]

Embedding model	ML Model	Accuracy	Precision	Recall	F1 score	EER score
wave2vec2	kNN	train 0.9442, dev 0.9262, eval 0.9104	train 0.9369, dev 0.9132, eval 0.6632	train 0.4837, dev 0.3097, eval 0.2693	train 0.6380, dev 0.4625, eval 0.3831	0.2258
wave2vec2	Random Forest	train 1.0000, dev 1.0000, eval 0.9233	train 1.0000, dev 1.0000, eval 0.8096	train 0.9996, dev 0.9996, eval 0.3365	train 0.9998, dev 0.9998, eval 0.4754	0.1266
wave2vec2	XGBoost	train 0.9564, dev 0.9498, eval 0.9308	train 0.9361, dev 0.9345, eval 0.7773	train 0.6132, dev 0.5487, eval 0.4617	train 0.6914, dev 0.7410, eval 0.5793	**0.1024**
wave2vec2	Naive Bays	train 0.7481, dev 0.7569, eval 0.8707	train 0.2655, dev 0.2771, eval 0.4361	train 0.8368, dev 0.8516, eval 0.8621	train 0.4031, dev 0.4181, eval 0.5792	0.1327
wave2vec2	Decision Tree	train 0.9839, dev 0.9820, eval 0.9042	train 0.9351, dev 0.9409, eval 0.5442	train 0.9043, dev 0.8803, eval 0.4453	train 0.9194, dev 0.9096, eval 0.4898	0.2483
wave2vec2	Log Regression	train 0.9326, dev 0.9267, eval 0.9341	train 0.7354, dev 0.7119, eval 0.7477	train 0.5267, dev 0.4800, eval 0.5459	train 0.6138, dev 0.5734, eval 0.6310	0.1141
wave2vec2	SVM	train 0.9483, dev 0.9417, eval 0.9372	train 0.8372, dev 0.8328, eval 0.7493	train 0.6101, dev 0.5396, eval 0.5884	train 0.7058, dev 0.6549, eval 0.6592	0.1171
wave2vec2	MLP	train 0.9939, dev 0.9913, eval 0.9316	train 0.9939, dev 0.9936, eval 0.6523	train 0.9461, dev 0.9207, eval 0.7222	train 0.9694, dev 0.9558, eval 0.6855	0.1044

6 Acknowledgement

I extend my sincere gratitude to Md Sahidullah, Assistant Professor at the Institute of Advancing Intelligence of TCG CREST, for his invaluable supervision and guidance throughout the duration of this project. His expertise, unwavering support, and insightful feedback have played a pivotal role in shaping the course of this work. I am truly thankful for the opportunity to benefit from his mentorship, which has significantly enriched my understanding and contributed to the success of this endeavor.

References

[1] Tomi Kinnunen, Kong Aik Lee, Hector Delgado, Nicholas Evans, Massimiliano Todisco, Md Sahidullah, Junichi Yamagishi, and Douglas A. Reynolds. t-dcf: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification, 2019.

[1] GitHub link of this project

[2] Amir Rostami, Mahdi Homayoonpoor, and Ahmad Nickabadi. Efficient attention branch network with combined loss function for automatic speaker verification spoof detection. *Circuits, Systems, and Signal Processing*, 42:1–19, 02 2023.

[3] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Hector Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sebastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-Francois Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, and Zhen-Hua Ling. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech, 2020.

[4] Jee weon Jung, Hee-Soo Heo, Hemlata Tak, Hye jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks, 2021.

[5] Junxiao Xue, Hao Zhou, Huawei Song, Bin Wu, and Lei Shi. Cross-modal information fusion for voice spoofing detection. *Speech Communication*, 147:41–50, 2023.