Rapport RDI

Détection d'usurpation d'identité dans l'audio

Noé Audemard

January 4, 2024

Contents

1 Abstract **1**

2 Introduction to voice spoofing **1**
2.1 Dataset 1

3 Voice Spoofing SOTA **2**
3.1 Results 2
3.2 Improved Model 3

4 Experimental setup for SASV **3**
4.1 Dataset 3
4.1.1 Training set 4
4.1.2 Testing set 4
4.2 Model 5
4.2.1 Naive modification 5
4.3 Training 5
4.4 Results and analysis 5

5 Informed modification **6**
5.1 Results and analysis 6

6 Summary and discussions **7**

7 **7**
Bibliography 7

1 Abstract

Voice spoofing, the act of imitating or falsifying someone's voice, has become a significant concern in various security-sensitive applications, including speaker verification systems, voice assistants etc. as well as in the context of misinformation, such as on social medias.As voice-based authentication methods gain popularity, the vulnerability to voice spoofing attacks poses a significant threat to the integrity and security of these systems.

This document provides an introduction to the current state of the art in spoofing detection. It then describes the chosen angle of research, the analysis of ECAPA-TDNN's resistance to a subset of voice spoofing attacks, voice conversion methods. It then studies a proposed change to its architecture to improve its abilities as a spoofing aware speaker verification model.

2 Introduction to voice spoofing

Voice spoofing methods can be classified in 3 majors categories: Physical attacks, logical attacks and impersonation attacks. The latter consists of a mimic, or someone with a very similar voice copying the targets. Physical attacks consist of using a recording of the target and replaying it to a microphone. Finally, logical attacks consist of creating an speech signal with a computer program and directly inputting it to the attacked system. Logical attacks can be generated by two types of methods: Voice synthesis, the process of creating an audio from a transcript and a speaker identity and voice conversion, the process of transferring the speaker identity from one audio to another.
It is possible to combine multiple types of attacks (Logical and physical for example), creating what is called 'Multi-order attacks'.

2.1 Dataset

ASVSpoof[9] is a competition that comes with a dataset that is the most commonly used for voice spoofing detection. ASVSpoof began in 2015 and has since been held every two years. The most recent version, 2021[5], contains 3 sections. Physical attacks, logical attacks, and deepfake attacks.
The logical attack section consists of genuine audios and audios created by TTS programs or voice conversion models, which are then transmitted

through VoIP systems to emulate telephony communication. The training set includes 6 spoofing methods, and the testing set contains 12. This means that the model's performance is tested on unseen attacks.
Deepfake attacks are similar to logical attacks, but the audios are encoded and decoded with different lossy codecs, similar to what is used as compression on television or social media. This section also contains a lot more spoofing methods than the logical section, with over 100 methods included. The repartition of those methods in the training and testing split is not disclosed.

3 Voice Spoofing SOTA

The best performing single model on ASVSpoof 2021 is the one proposed by 'Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation'[10]. The model used in this paper is an improvement of 'End-to-End Spectro-Temporal Graph Attention Networks for Speaker Verification Anti-Spoofing and Speech Deepfake Detection' [7].
This model is end to end, meaning that it takes raw waveform as input. It uses a sinc convolution layer to extract learnable higher level feature map. It then splits up, with two similar blocks in parallel. Both are composed of a single GAT layer and an attention block. One of the blocks has a temporal attention layer while the other has a spatial attention layer. Then, a graph pooling layer is used to combine the outputs of both. This layer has 3 variants based on the way it merges the inputs: Addition, Multiplication and concatenation.
After this layer, a final block containing a GAT layer and an spatial and temporal attention block is then projected to get the predicted classes.

3.1 Results

In order to verify the results of this paper, the evaluation of its results was recomputed for all 3 given versions of the model and the best performing model was retrained, albeit with a lower number of epochs. The results are in table 1.

	Paper results	Provided model	Retrained model
RawGatST Add	1.15	1.15	- -
RawGatST Mutliply	**1.06**	**1.06**	1.33
RawGatST Concat	1.23	1.23	- -

Table 1: EER% on ASVSpoof 2019 evaluation set of RawGatST

3.2 Improved Model

The improved model proposed in [10] has 2 additions:
- A better data augmentation
- A new feature extractor
The new data augmentation is in addition to the one used in the original paper.
The new version also uses 'RawBoost'[8].
The new feature extractor is a pretrained version of 'Wav2Vec 2.0'[4], a self-supervised speech recognition model
. This new backend extractor leverages the larger non-supervised speech datasets available to produce a more robust speaker identity representation. The combination of the 2 changes improves the EER% on ASVSpoof2019 LA from 1.06 to 0.82, a 20% relative improvement.

4 Experimental setup for SASV

The following sections will describe the reflexion process and choices made in order to evaluate Spoofing Aware Speaker Verification (SASV) systems based on current Speaker Verification models.

4.1 Dataset

Since ASVSpoof does not reference the original audios and speakers used for each spoofed audios in its database, it is not suitable for training a spoofing aware automatic speaker verification system.
In order to train such a model, we will need to create our own base of spoofed audios. In order to limit the computation time necessary to create this database, we will limit ourselves to voice conversion methods, as they are

much faster to compute than voice synthesis methods.

4.1.1 Training set

The chosen voice conversion method for this dataset is 'AgainVC' [6]. This model uses an encoder-decoder architecture with the encoder producing two separatete outputs: A speaker identity encoder and a context encoder. This method leverages instance normalization and activation guidance to separate the information in the two embeddings.
We also need an original database of audios to serve as sources for the converted audios and to be the genuine audios of the new database. We choose 'Voxceleb1'[1].
We generate 3 spoofed audios for each of the 1251 speakers in Voxceleb1. This way, each speaker has 3 audios converted to its speaker identity, from 3 random other speakers. We split those spoofed audios by destination speaker according to the training and testing set of Voxceleb1.

4.1.2 Testing set

In order to evaluate the performance of the SASV on unseen attacks, we also need to create a dataset with a different method. We are looking for another voice conversion method that matches the following requirements: - Pretrained model available - Different core architecture to AgainVC. This is important to test the model's ability to generalize - Better performance than AgainVC. This is necessary to make sure our SASV is robust against newer and improved attacks, which is necessary for real world use cases. The combination of those requirements led me to choose CycleGAN-VC2[2] as the voice conversion method used in the testing dataset. The fundamental concept of CycleGAN-VC2 is to utilise an adversarial training approach, which involves the use of two generators and two discriminators. The first generator is responsible for converting a source voice to a target voice, while the second generator reconverts the converted voice back to the original source voice. The discriminators, on the other hand, aim to distinguish between real and generated voices. This cycle-consistent training approach helps to ensure that the generated voices maintain the linguistic content and speaker identity of the original voices. CycleGAN-VC2 also uses an MRF (Multi-Receptive Field fusion module) to capture both local and global features.

4.2 Model

We choose to train a speaker verification model for this task to evaluate its robustness to spoofing attacks and understand if classical speaker verification architectures can be adapted to resist spoofing attacks.
In this endeavor, we choose to train ECAPA-TDNN[3] since it is on of the best performing simple single models for speaker verification.

4.2.1 Naive modification

In order to leverage both the information on speaker identity and genuine/spoofed audios, we need to adapt the model. In order to evaluate the performances of the model with the least modifications, the chosen change is to add a new class to the model output. Since ECAPA-TDNN is a classification model which learns one class for each speaker (in the training set), we will add one class and label all spoofed audios as belonging to this class. This means that we can keep the existing loss and minimize the change between the original ECAPA-TDNN and our SASV.

4.3 Training

We initially train the model on only the training set of Voxceleb1 and evaluate its performance on the test set of Voxceleb and of our spoofed audios.
We then train the model on both the training set of Voxceleb1 and of our generated audios and evaluate the performance on both test sets. Finally we test both the models on the audios generated with CycleGAN-VC2. One of the main enhancements of CycleGAN-VC2 is the use of a multi-receptive field fusion (MRF) module

4.4 Results and analysis

	Genuine audios	**Again-VC**	**CycleGAN-VC2**
Trained on genuine audios	3.39	26.4	35.1
Trained on both	4.81	20.5	33.2

Table 2: EER% on Genuine audios, Again-VC audios and CycleGAN-VC2 audios with naive ECAPA-TDNN based model

We can start by looking at the first column. We notice that training the model on more than just the genuine audios reduces its performance significantly from 3.39% EER to 4.81% EER. For the second column, we can see that the initial EER% is high but still far from 50%, which shows that even the model trained only on speaker verification can correctly reject spoofed audios. It is important to note that this might be in part due to the very low quality of some of the audios produced by Again-VC. We can also see that the EER% decreases significantly when the model is also trained on Again-VC spoofed audios. This confirms that the ECAPA-TDNN architecture is able to learn to discriminate some spoofed audios, but the EER% remains significantly higher than dedicated speaker verification models. Finally, the third column shows that training on one spoofing method does increase performances on different, more advanced methods, even though the relative change in EER% is much lower that on the methods it was directly trained on.

5 Informed modification

We would now like to see if a less naive modification of ECAPA-TDNN allows for better performances. The major problem with the initial modification was that it didn't keep information on the initial class of the voice speaker. In order to solve this, we will keep the initial multiclass output and loss and add a binary classifier with BCE loss (binary cross entropy). The final computed loss is the sum of the initial loss and of the BCE multiplied by a constant (Hyperparameter). We train and test this model with the same setup as the naive one.

5.1 Results and analysis

	Genuine audios	**Again-VC**	**CycleGAN-VC2**
Trained on genuine audios	3.32	25.9	34.4
Trained on both	4.89	21.2	33.0

Table 3: EER% on Genuine audios, Again-VC audios and CycleGAN-VC2 audios with naive ECAPA-TDNN based model

We can see that the results are very similar to the ones of the naive model. Some EER% are higher than others while some are lower, but they are all too close to deduce statistical significance. The proposed improvement does not seem to improve performances, even though a larger and more diverse test set would be necessary for a conclusive result.

6 Summary and discussions

Voice spoofing detection is very important to preserve the trust in the voice we hear online and on the phone, as well as to protect speaker verification systems. With this goal, adapting existing speaker verification models to protect against spoofing attacks is crucial. We proposed an experimental setup to evaluate the feasibility of this method on current speaker verification model architecture ECAPA-TDNN. We showed that simple modifications to the model to allow training on spoofed audios lets the model learn to discriminate spoofed audios. The results also showed that such a model is not able to reach performances similar to dedicated spoofing detection models coupled with classical speaker verification systems.

7

Bibliography

[1] Arsha Nagrani Joon Son Chung Andrew Zisserman. "VoxCeleb: a large-scale speaker identification dataset". In: (2017).

[2] Takuhiro Kaneko Hirokazu Kameoka Kou Tanaka Nobukatsu Hojo. "CYCLEGAN-VC2: IMPROVED CYCLEGAN-BASED NON-PARALLEL VOICE CONVERSION". In: (2019).

[3] Brecht Desplanques Jenthe Thienpondt Kris Demuynck. "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification". In: (2020).

[4] Alexei Baevski Henry Zhou Abdelrahman Mohamed Michael Auli. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: (2021).

[5] Xuechen Liu Xin Wang Md Sahidullah Jose Patino Hector Delgado Tomi Kinnunen Massimiliano Todisco Junichi Yamagishi Nicholas Evans Andreas Nautsch Kong Aik Lee. "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild". In: (2021).

[6] Yen-Hao Chen Da-Yi Wu Tsung-Han Wu Hung-yi Lee. "AGAIN-VC: A ONE-SHOT VOICE CONVERSION USING ACTIVATION GUIDANCE AND ADAPTIVE INSTANCE NORMALIZATION". In: (2021).

[7] Hemlata Tak Jee-weon Jung Jose Patino Madhu Kamble Massimiliano Todisco and Nicholas Evans. "End-to-End Spectro-Temporal Graph Attention Networks for Speaker Verification Anti-Spoofing and Speech Deepfake Detection". In: (2021).

[8] Hemlata Tak Madhu Kamble Jose Patino Massimiliano Todisco and Nicholas Evans. "RAWBOOST: A RAW DATA BOOSTING AND AUGMENTATION METHOD APPLIED TO AUTOMATIC SPEAKER VERIFICATION ANTI-SPOOFING". In: (2021).

[9] Hsin-Te Hwangh Yu Tsaoh Hsin-Min Wangh Sebastien Le Magueri Markus Beckerj Fergus Hendersonj Rob Clarkj Yu Zhangj Quan Wangj Ye Jiaj Kai Onumak Koji Mushikak Takashi Kanedak Yuan Jiangl Li-Juan Liul Yi-Chiao Wum Wen-Chin Huangm Tomoki Todam Kou Tanakan Hirokazu Kameokan Ingmar Steinero Driss Matroufp Jean-Francois Bonastrep Avashna Govenderb Srikanth Ronankiq Jing-Xuan Zhangr Zhen-Hua Ling Xin Wanga Junichi Yamagishia Massimiliano Todiscoc Hector Delgadoc Andreas Nautschc Nicholas Evansc Md Sahidullahd Ville Vestmane Tomi Kinnunene Kong Aik Leef Lauri Juvelag Paavo Alkug Yu-Huai Pengh. "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed". In: (2021).

[10] Hemlata Tak Massimiliano Todisco Xin Wang Jee-weon Jung Junichi Yamagishi and Nicholas Evans. "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation". In: (2021).