Malafide: a novel adversarial convolutive noise attack against deepfake and spoofing detection systems

Michele Panariello*, Wanying Ge*, Hemlata Tak, Massimiliano Todisco and Nicholas Evans

EURECOM, Sophia Antipolis, France
firstname.lastname@eurecom.fr

1. Introduction

Spoofing, or presentation attacks can be used by a fraudster to manipulate the behaviour of a biometric recognition system and hence to gain illegitimate access to protected systems, services or facilities. Auxiliary sub-systems in the form of countermeasures (CMs) are nowadays commonly deployed in order to defend against such attacks and can offer strong levels of protection [1, 2], including in the case of automatic speaker verification (ASV) [3, 4], the focus in this paper. CMs aim to detect tell-tale signs of spoofing, namely processing artefacts that are not expected in recordings of bona fide, genuine human speech.

Spoofing and detection research are nonetheless a game of cat and mouse in which a defender continually adapts to emerging threats, while an attacker or adversary continually adapts to the resulting CMs. While threats to ASV have evolved in the form of ever-more-effective synthetic speech and converted voice attacks, CMs have largely kept apace [5–7]. New threats nonetheless continue to emerge.

With defences now commonplace, adversaries can adapt to conceal the artefacts which might otherwise serve to distinguish bona fide from spoofed biometric samples; they can devise new attacks to manipulate not just the biometric classifier, but also a spoofing CM. Examples of such adversarial attacks have already emerged [8–13]. Most take the form of additive noise whereby utterance-specific perturbations are learned and added to a speech signal. These approaches cannot be implemented in real time and are sensitive to the specific utterance; the perturbations tend to be highly sensitive and even slight distortions can render the attack ineffective [14, 15].

* These authors contributed equally to this work.

In this paper, we propose Malafide, a novel attack designed for the strength-testing of spoofing and deepfake detection solutions against adversarial, convolutive noise attacks. Malafide attacks involve the optimisation of a linear time-invariant filter which is applied to deepfake or spoofed utterances in order to provoke their misclassification as bona fide utterances. Convolutional noise is independent of an utterance and its duration and is naturally robust to time-domain shifts, unlike additive adversarial noise. Malafide attacks require the optimisation of only a small number of filter coefficients, far less than the number of waveform samples or, equivalently, the number of samples that would need to be generated in an additive noise attack. They are also more universal than additive noise attacks in that they are not optimised for each utterance or speaker, but are instead optimised for a given spoofing attack, with the latter acting to ensure they compromise both CM and ASV subsystems. The attack is optimised offline and can hence be applied in real time.

While attack studies of the nature discussed above can raise obvious ethical questions, such work is key to the typical adversarial development cycle. Only by continually probing and strength-testing a given system and by addressing any identified weaknesses can there be any confidence in its security. This is the spirit of our work presented in this paper.

2. Relation to prior work

Adversarial attacks, first introduced for image related tasks [16], have also been studied in the speech domain, e.g. for automatic speech recognition (ASR) [14, 17] as well as spoofing and ASV [8, 9, 11–13, 15]. Early studies [8, 16] explored adversarial examples in the form of additive noise and attack transferability in black-box scenarios. Typically, these attacks operate on feature representations, implying that the attacker has corresponding system-level access. This is unlikely in practice and the attack threat diminishes greatly if time-domain speech signals are resynthesized from features [12]. Malafide attacks are applied at the level of raw speech signals in the time domain.

The early approaches consider the generation of utterance-specific adversarial noise. Again drawing upon inspiration from studies in image processing [18], universal adversarial perturbations have been adapted to a wide variety of audio tasks, such as ASR [19], ASV [20], and environmental sound classification [21]. Common to these approaches is the generation of adversarial perturbations by iteratively optimising over several data points. We adopt a similar approach, but generate adversarial noise that is specific to an underlying spoofing attack.

To the best of our knowledge, the most relevant prior work is [11], which reports an investigation of universal perturbations against spoofing and deepfake CM systems. This technique, though, targets the simultaneous manipulation of both
CM and ASV subsystems in a manner that is independent to specific spoofing attacks. So that they compromise the ASV system, perturbations are also generated for specific speakers, which adds complexity and precludes their usage on unseen speakers. Moreover, there is no explicit constraint that acts to protect speech quality. Our approach differs in that it operates alongside specific spoofing attacks to augment the threat they pose to combined CM and ASV subsystems. We assume that the spoofing attack is sufficient to manipulate the ASV subsystem but, by protecting other speech attributes such as intelligibility, prosody and the speaker’s voice, Malafide attacks act to compromise both CM and ASV subsystems. Different to all previous work, our approach involves the learning of an adversarial linear, time-invariant (LTI) filter which can be applied in real-time to a spoofed utterance through time domain convolution.

3. Malafide attacks

Let \( S^{(a)} = \{s_1^{(a)}, \ldots, s_N^{(a)}\} \) be a set of deepfake/spoofed utterances generated by algorithm \( a \) (a particular text-to-speech or voice conversion algorithm). Spoofed utterances are generated to manipulate an ASV system so as to increase the likelihood of it verifying erroneously claimed identities. Auxiliary detection classifiers in the form of CMs are used to defend against spoofing attacks and hence to protect ASV reliability. Let \( CM(u) = s(y | u) \) be a model that assigns a score \( y \) to utterance \( u \) where, by convention, higher scores reflect greater support for the bona fide class and lower scores greater support for the spoof class. Ideally, for most spoofed utterances \( i \), \( CM(s_i^{(a)}) \) will produce low scores.

Malafide attacks involve the optimisation of a linear time-invariant (LTI), non-causal filter, the coefficients (tap weights) of which are optimised to provoke the misclassification of deepfake/spoofed utterances as bona fide utterances. The LTI, \( L \)-tap filter \( m^{(a)}(t) \) is designed to maximise \( CM(s_i^{(a)} * m^{(a)}) \) (where * denotes convolution). In the case of several different spoofing algorithms \( a_1 \ldots a_K \), an attacker can optimise an equivalent number of filters \( m_1^{(a_1)} \ldots m_K^{(a_K)} \). The learning of attack-specific filters is intuitive given that spoofing artefacts are also attack-specific. The filter should then be tuned to counter the reliance of the CM upon these same artefacts.

Filter coefficients \( m^{(a)} \) can be optimised with conventional gradient descent using the set of spoofed utterances \( S^{(a)} \). The objective function is given by

\[
\max_{m^{(a)}} \sum \ CM(s_i^{(a)} * m^{(a)})
\]

A graphical depiction of the training procedure is shown in Figure 1.
4. Experimental setup

4.1. Protocols and filter optimisation

All experiments were conducted using the ASVspoof 2019 logical access (LA) dataset [23]. It contains spoofing attacks generated with a set of algorithms labelled A01 to A19. Attacks A01 to A06 are contained in both the training and development partitions, while A07 to A19 are contained only in the evaluation partition. Training and development partitions relate to the realm of a defender whose role is to train and develop spoofing CMs. The set of three CMs described in Section 4.3 are trained in the usual way using these two data partitions.

In contrast, the test partition contains data in the realm of the attacker. Attack-specific filters are hence trained according to (1) using a subset of the test partition, i.e. using A7 to A19 spoofing attack data. We stress that, in contrast to usual practice, the use of test data for training purposes is acceptable in this case; the attacker is not bound by experimental protocols and can use test data in any reasonable way that is to their advantage. The test partition is nonetheless split into two parts which contain an equal number of utterances for each attack. Attack-specific adversarial filters are then optimised using only data in Part 1 and tested using data in Part 2.

This re-partitioning allows us to verify the universality of Malafide attacks, namely the effectiveness or transferability of the attack to unseen utterances. The setup reflects a scenario in which filters are trained by an attack offline and then used to implement online/real-time attacks, e.g. in a logical access or telephony scenario. While all data partitions contain both bona fide and spoofed data, Malafide filter optimisation is performed using spoofed utterances only.

4.2. Implementation

(1) is optimized with Adam [24]. The learning rate and weight decay are tuned separately for each CM model. Filters are optimised for 15 epochs with a batch size of 14 using data in Part 1 (see Section 4.1). During optimisation of (1), the weights of the CM model are kept frozen. We explored different filter lengths $L$ (65, 129, 257, 513, 1025, 2049 and 4097) in order to explore the balance between optimisation of (1) and the preservation of speech fidelity (see Section 3).

The filter used for evaluation is selected according to a measure of the attack success rate for the full set of utterances in Part 1. The attack success rate is defined as the fraction of spoofed utterances for which $N(CM(s \cdot m)) > 0.5$, where $N(\cdot)$ normalises CM scores to reflect probabilities in $[0, 1]$. This setting reflects the point at which the CM considers $s$ more likely to be bona fide than spoofed. Our specific implementation is available as open-source and can be used to reproduce our results under the same GPU environment.

4.3. Countermeasures

We used three different CM systems to verify the effectiveness of the adversarial filter attack. They are described below. All are available as open-source.

RawNet2 [25], an end-to-end (E2E) model developed originally for ASV, has also been applied to spoofing and deepfake detection [26]. The first network layer is a bank of 20 mel-scaled sinc filters, and is convolved directly with raw waveform inputs. The sinc-layer is followed by a series of residual blocks and a gated recurrent unit (GRU) which produces a score indicative of whether the input is bona fide or spoofed.

AASt [7] is a state-of-the-art E2E spoofing CM solution based upon the RawNet2 CM described above. It uses the same sinc-layer and residual network to extract higher-level feature representations. The back-end includes a spectro-temporal graph attention network (RawGAT-ST) [27], heterogeneous graph attention layers and max graph operations to integrate temporal and spectral representations. Scores are generated using a readout operation and a fully connected output layer.

Self-supervised learning (SSL) based front-ends have gained increasing attention in a range of speech-related tasks in recent years, including for spoofing and deepfake detection [28–31]. The SSL-based CM architecture [6] is a two-stage model with SSL-based feature extraction and a back-end comprising graph attention and pooling layers, a single fully-connected layer and an output layer. It is the only of the three CMs used in this work, which is trained using external data beyond that permitted by ASVspoof evaluation rules. Nonetheless, SSL solutions have improved substantially on the previous state of the art [5,6]. The SSL feature extractor is a pre-trained wav2vec 2.0 model [32] with the weights of which are fine-tuned during CM training.

4.4. Metrics

All results are reported as equal error rate (EER) estimates and are obtained using the standard SASV evaluation protocol [4]. Results reported in Section 5 are CM EERs computed using a mix of bona fide and spoofed trials. Results reported in Section 6 are SASV-EERs computed using a mix of target (positive class) and both non-target and spoofed utterances (negative class).

5. Results

Results are presented in Table 1. They show EERs without and with the use of different length adversarial filters (column 1). Results are also shown for filters optimised for one CM (row 1) and tested against another (row 2); column 2 shows EERs for a filter optimised for, and tested with the AASt CM (a white-box setting), whereas column 3 shows EERs for a filter optimised for the AASt CM but tested with the RawNet2 CM (a black-box setting).

Results for the three white-box settings shown in columns 2, 6 and 10 of Table 1 show that adversarial filtering provokes substantial increases in the CM EER, with the greatest EERs being achieved with filters of either 513 coefficients (AASt and RawNet2) or 1025 coefficients (SSL). With a maximum EER of 12.3%, the SSL CM is the most robust. The most vulnerable is the RawNet2 CM for which the maximum EER is 22.0%, although it also has the highest initial EER without adversarial filtering of 3.3%.

We now turn to black-box settings. Columns 3 and 4 of Table 1 show black-box results for filters learned using the AASt CM. The attack is transferable to the RawNet2 CM (maximum EER of 23.9%), but less so to the SSL CM (maximum...
Table 1: CM performance in terms of the EER (%) in both white-box and black-box settings. Results shown without filtering and with Malafide filtering using filters of different lengths L (number of taps).

<table>
<thead>
<tr>
<th>Filter length</th>
<th>AASIST Malafide Attack</th>
<th>RawNet2 Malafide Attack</th>
<th>SSL Malafide Attack</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(white-box)</td>
<td>(black-box)</td>
<td>(black-box)</td>
</tr>
<tr>
<td></td>
<td>(black-box)</td>
<td>(black-box)</td>
<td>(black-box)</td>
</tr>
<tr>
<td></td>
<td>AASIST</td>
<td>RawNet2</td>
<td>SSL</td>
</tr>
<tr>
<td>no filter</td>
<td>0.71</td>
<td>3.29</td>
<td>1.01</td>
</tr>
<tr>
<td>65</td>
<td>5.54</td>
<td>8.94</td>
<td>3.63</td>
</tr>
<tr>
<td>129</td>
<td>8.15</td>
<td>10.21</td>
<td>1.76</td>
</tr>
<tr>
<td>257</td>
<td>9.73</td>
<td>9.72</td>
<td>0.97</td>
</tr>
<tr>
<td>513</td>
<td>13.87</td>
<td>11.18</td>
<td>0.19</td>
</tr>
<tr>
<td>1025</td>
<td>12.71</td>
<td>15.81</td>
<td>0.15</td>
</tr>
<tr>
<td>2049</td>
<td>9.36</td>
<td>23.93</td>
<td>0.26</td>
</tr>
<tr>
<td>4097</td>
<td>6.62</td>
<td>19.18</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>1.27</td>
<td>11.92</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Informal listening tests revealed that shorter-length filters better preserve speech fidelity whereas longer-length filters result in detectable reverberation. Longer-length filters degrade speech fidelity to the point that spoofing attacks are no longer successful in compromising the ASV system.

We acknowledge that SASV-EERs are heavily dependent on the proportion of negative class trials that are spoofed (as opposed to non-target). In this respect, performance estimates are not necessarily indicative of what might be expected in the wild where, for instance, spoof attacks may be less prevalent.

7. Conclusions

The work reported in this paper shows that the reliability of spoofing countermeasures (CMs) can be compromised using adversarial, linear time-invariant filters and that these can also be configured to compromise integrated CM and automatic speaker verification (ASV) systems. Malafide attacks are a threat in both white-box and black-box settings, and a RawNet2 CM is particularly vulnerable.

Results for integrated CM and ASV systems show that Malafide attacks are successful in manipulating a spoofing CM and, when used in conjunction with spoofing attacks and by introducing only modest perturbations, ASV subsystems too. The performance of the integrated system that uses a self-supervised learning (SSL) CM is an exception; performance is reasonably robust. This is likely caused by the introduction of greater distortion to the speech signal that is needed to compromise the relatively more complex CM. The same distortion interferes with the ability of the spoofing attack to compromise the ASV system. Future work should study similar convolutional attacks that are optimised to compromise both CM and ASV subsystems. These attacks might expose vulnerabilities of even SSL-based systems. Such work and is critical if we are to protect confidence in the reliability of voice biometrics technology.

8. Acknowledgements

The second author is supported by the TReSPAsS-ETN project funded by the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860813. The third author is supported by the VoicePersonae project funded by the French Agence Nationale de la Recherche (ANR) and the Japan Science and Technology Agency (JST).
9. References


