The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)
31 August 2024, Kos, Greece

Malacopula: adversarial automatic speaker verification attacks using a neural-based generalised Hammerstein model

Massimiliano Todisco[1], *Michele Panariello*[1], *Xin Wang*[2],
Héctor Delgado[3], *Kong Aik Lee*[4], *Nicholas Evans*[1]

[1]EURECOM, Sophia Antipolis, France, [2]National Institute of Informatics, Japan
[3]Microsoft, Spain, [4]The Hong Kong Polytechnic University, Hong Kong
[1]*firstname.lastname*@eurecom.fr

Abstract

We present Malacopula, a neural-based generalised Hammerstein model designed to introduce adversarial perturbations to spoofed speech utterances so that they better deceive automatic speaker verification (ASV) systems. Using non-linear processes to modify speech utterances, Malacopula enhances the effectiveness of spoofing attacks. The model comprises parallel branches of polynomial functions followed by linear time-invariant filters. The adversarial optimisation procedure acts to minimise the cosine distance between speaker embeddings extracted from spoofed and bona fide utterances. Experiments, performed using three recent ASV systems and the ASVspoof 2019 dataset, show that Malacopula increases vulnerabilities by a substantial margin. However, speech quality is reduced and attacks can be detected effectively under controlled conditions. The findings emphasise the need to identify new vulnerabilities and design defences to protect ASV systems from adversarial attacks in the wild.

1. Introduction

The performance of automatic speaker verification (ASV) systems has improved remarkably in recent years. The pioneering x-vector approach [1] laid the foundation for more recent and robust systems including ECAPA [2], CAM++ [3], and ERes2Net [4] which consistently outperform their predecessors in various benchmarks [5].

Despite these technological advances, ASV systems remain vulnerable to spoofing attacks implemented using, e.g., text-to-speech synthesis and voice conversion techniques. These attacks have become increasingly sophisticated, capable of producing spoofed speech which is generally indistinguishable from bona fide speech and which effectively compromise ASV reliability. Nonetheless, there is evidence that recent ASV systems have some natural defensive capabilities against spoofing attacks [6]. Natural defences can also be supplemented using auxiliary spoofing and deepfake detection solutions [7].

While the study of spoofing and the development of detection solutions has attracted broad attention, a new threat has emerged in the form of adversarial attacks, e.g. [8]. These are implemented using adversarial training which, in the context of ASV, can be used by a fraudster to introduce noise—sometimes imperceptible or easily mistakable for real environmental sounds—to deceive the classifier and provoke a higher rate of false alarms/acceptances.

In recent work [9], we showed how adversarial training techniques can be used to design a simple linear time-invariant (LTI) filter, named Malafide, which compromises the reliability of even state-of-the-art spoofing and deepfake detection solutions. In this paper we report our work to evaluate the robustness of ASV systems to the same form of adversarial attacks. We introduce Malacopula,[1] a neural-based generalised Hammerstein model [10] designed specifically to compromise ASV system reliability through the introduction of adversarial perturbations to a test speech utterance. Unlike Malafide, Malacopula supports the modification of not only amplitude and phase but also frequency components in non-linear fashion, a crucial benefit for voice cloning.

Malacpolua acts as a post-processing filter to increase ASV system vulnerabilities to spoofing attacks. Tuned to the spoofing attack and speaker identity, the Malacopula filter is optimised independently of the utterance and input duration, requiring the optimisation of only a small number of filter coefficients, in similar fashion to Malafide [9].

2. Literature Review

Adversarial attacks were originally introduced for image processing tasks [11], but have since been applied to the speech domain, particularly focusing on automatic speech recognition (ASR) [8, 12] and spoofing/automatic speaker verification (ASV) systems [13–15].

Early strategies involved generating adversarial noise specific to each utterance, drawing inspiration from image processing techniques [16]. These strategies adapted universal adversarial perturbations to various audio tasks, including automatic speech and speaker recognition [17–19]. A common theme among these methods is the iterative optimisation of adversarial perturbations across multiple data samples.

Initial research [11, 15] primarily explored adversarial examples as additive noise and their ability to transfer to unseen scenarios. The study in [14] explored universal perturbations against spoofing and deepfake countermeasure (CM) systems. This method targets both CM and ASV subsystems independently of specific attacks. However, it requires the generation of a different array of adversarial noise for each utterance, which results in a high computational effort. Moreover, the variable length of speech utterances is a constraint upon the generation of adversarial noise, rendering these attacks impractical in real-world scenarios.

Malafide [9] introduced an adversarial technique utilising linear time-invariant (LTI) filters applied in real-time to spoofed utterances through time-domain convolution. Unlike traditional

[1]*Mala copula* is Latin for "bad connection" or "bad union." It signifies an undesirable or improper association between elements.

10.21437/ASVspoof.2024-14

methods, Malafide filters are optimised independently of the input utterance and its duration, tailored specifically to the underlying spoofing attack. This method requires the optimisation of only a small number of filter coefficients, thereby offering greater flexibility in its application.

Our approach takes a different path by enhancing specific spoofing attacks and targeting specific speakers to increase the threat to ASV systems. We operate under the assumption that the spoofing attack effectively manipulates the ASV subsystem.

In contrast to prior research, our method involves the use of adversarial, non-linear filters using a generalised Hammerstein model, commonly used for the identification of non-linear systems. Malacopula can be applied in real-time to a spoofed utterance via time-domain convolution operations, specifically targeting a particular speaker and the underlying attack algorithm.

3. Generalised Hammerstein Model

The generalised Hammerstein model is a prominent framework in signal processing, employed to identify non-linear dynamic systems. The model combines a static non-linear component with a linear dynamic component, enabling detailed representation and manipulation of complex signal characteristics.

The structure of the Generalised Hammerstein Model typically comprises two main elements: a non-linear transformation followed by a linear time-invariant (LTI) filter. Mathematically, the model can be expressed as:

$$y[n] = \sum_{k=1}^{K} \sum_{i=0}^{L-1} h_k[i] \phi_k(x[n-i]) \tag{1}$$

where $y[n]$ is the output signal, $x[n]$ is the input signal, $\phi_k(x[n-i])$ represents the static non-linear transformation, $h_k[n]$ represents the impulse response of the LTI filter for the k-th branch, L denotes the memory length, K is the number of parallel branches and n represents the discrete sample index. The non-linear transformation captures the non-linearities of the input signal, often modelled using polynomials as functions of the input signal: $\phi_k(\cdot) = (\cdot)^k$. The versatility and computational efficiency of the generalised Hammerstein model have facilitated its successful application to the modelling of non-linear systems across various fields, including audio processing, acoustics, and mechanical vibrations [20–23].

Polynomial Hammerstein models have been employed to characterise and model non-linear loudspeakers using empirically measured Volterra kernels [24]. Results show the potential of the approach in estimating reliable non-linear models which accurately predict the response to complex real-speech inputs. For the collection of the ASVspoof 2019 physical access (PA) databases [25], the generalised Hammerstein model was utilised to model and simulate loudspeaker artefacts which often stem from non-linear behaviour. Both linear and non-linear characteristics accurately simulating the distortions introduced by loudspeakers. RawBoost [26] leverages the generalised Hammerstein model within a machine learning framework primarily for augmentation purposes rather than system identification. RawBoost simulates a wide range of signal distortions, thereby improving the robustness and generalisation of machine learning models trained with augmented data. RawBoost was developed specifically for the detection of spoofing and deepfakes in the wild but has also been used effectively in other speech-related applications [27].

4. Malacopula

The generalised Hammerstein model offers a powerful method to manipulate multiple characteristics of a speech signal, including the modification of amplitude and phase, but also frequency components in non-linear fashion. This capability can be exploited by malicious actors to create adversarial perturbations in order to deceive ASV systems. In the following we describe the implementation of the Malacopula filter.

4.1. Malacopula filter architecture

The Malacopula filter structure is shown in Fig. 1. It is composed of K parallel branches, which represent the non-linear depth, each involving a linear filter $\mathbf{c}$ of length L modulated by a Bartlett window $\mathbf{w}$.[2] Each branch processes the input signal $\mathbf{x}$ by a k-th non-linear, static power polynomial function. The filter operates entirely in the discrete time domain using convolution operations.

Mathematically, the filter is defined by:

$$mc_{K,L}(\mathbf{x}) = \sum_{k=1}^{K} \left[\mathbf{x}^k * (\mathbf{w} \odot \mathbf{c}_{k,L}) \right] \tag{2}$$

where $*$ denotes the convolution operator, and $\odot$ represents the Hadamard product.

Additionally, a normalisation layer using the L_∞ norm is applied after the summation operator to produce the output:

$$MC(\mathbf{x}) = \frac{mc(\mathbf{x})}{|mc(\mathbf{x})|_\infty} \tag{3}$$

4.2. Adversarial Optimisation Procedure

The Malacopula optimisation procedure is illustrated in Fig. 2. Each filter is trained independently for a given speaker s, and a spoof utterance $\mathbf{x}$ generated with spoofing algorithm a, and a bona fide enrolment utterance $\mathbf{y}$. The neural-based generalised Hammerstein model minimises the following objective function:

$$\min_{\mathbf{c}_{K,L}^{(s,a)}} \left[1 - CS\left(f_A\left(MC_{K,L}^{(s,a)}(\mathbf{x}) \right), f_A(\mathbf{y}) \right) \right] \tag{4}$$

where $f_A(\cdot)$ denotes the pre-trained speaker embedding extractor, and $CS(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$ is the cosine similarity between embeddings $\mathbf{A}$ and $\mathbf{B}$. The objective function aims to minimise the cosine distance between the speaker embeddings of the modified input signal $MC(\mathbf{x})$ and the target signal $\mathbf{y}$, ensuring that the adversarial perturbations are effective.

To ensure that the adversarial attacks generalise well across different speaker embeddings, the best Malacopula filter is selected using another speaker embedding extractor $f_B(\cdot)$. Filter selection is based on the minimum Wasserstein distance[3] computed across all training iterations, and between the following

[2] A Bartlett window, also known as a triangular window, is used in signal processing to reduce the side lobes of the filter response, which helps in minimising the spectral leakage. This window is chosen because it provides a trade-off between the width of the main lobe and the level of side lobes, making it suitable for applications where both frequency resolution and dynamic range are important.

[3] The Wasserstein distance is chosen because it provides a robust measure of the similarity between two probability distributions by considering the 'cost' of transforming one distribution into another. This property is particularly useful in evaluating the similarity between bona fide and spoof scores, as it captures differences in both the shape and location of the distributions, ensuring that adversarial examples remain

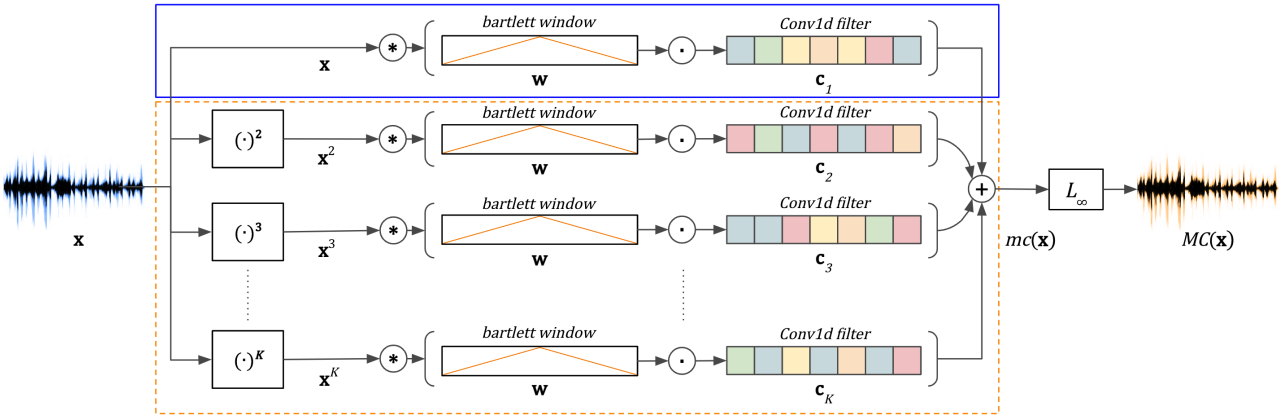

Figure 1: Malacopula filter architecture based on the generalised Hammerstein model. The blue box represents the linear component, while the the orange dashed box represents the non-linear filter components.

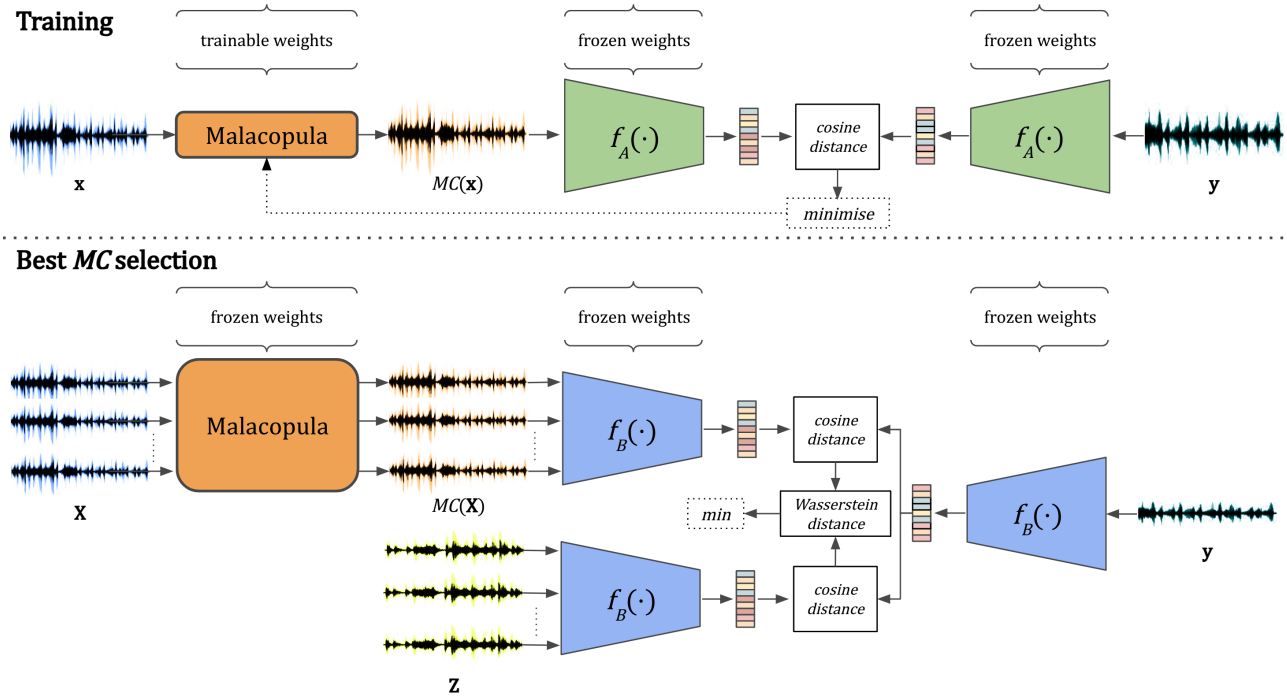

Figure 2: During training, Malacopula filters are optimised with the speaker embedding extractor $f_A(\cdot)$ as denoted by Equation 4. To ensure generalisation across different speakers, the best Malacopula filter is selected using another speaker embedding extractor $f_B(\cdot)$. The selection is based on the minimum Wasserstein distance between the following two score distributions: (i) the cosine distance between spoofed utterances processed by the Malacopula filter $MC(\mathbf{X})$ and bona fide enrolment utterances $\mathbf{y}$, and (ii) the cosine distance between bona fide target utterances $\mathbf{Z}$ and bona fide enrolment utterances $\mathbf{y}$. If multiple enrolment utterances are available, we use the average enrolment embedding.

two score distributions: (i) the cosine distance between embeddings extracted from spoofed utterances processed by the Malacopula $MC(\mathbf{X})$ and those extracted from bona fide enrolment utterances $\mathbf{y}$, and (ii) the cosine distance between embeddings extracted from bona fide target utterances $\mathbf{Z}$ and bona fide enrolment utterances $\mathbf{y}$. If multiple enrolment utterances are available, we use the average of their embeddings as the final enrolment embedding. Here $\mathbf{X}$ and $\mathbf{Z}$ are batches of speech utterances. This approach hence ensures that adversarial examples are sufficiently similar to the voice of the original speaker. Specifically, we employ a signed Wasserstein distance to incorporate not only the magnitude but also the direction of the distance. A positive Wasserstein distance is considered if the me-

close to bona fide data. Unlike the Equal Error Rate (EER), which only considers the point where the false acceptance rate equals the false rejection rate, the Wasserstein distance evaluates the entire distribution, offering a more comprehensive assessment of distribution similarity. The EER may not effectively capture the nuances of distributional changes caused by adversarial perturbations.

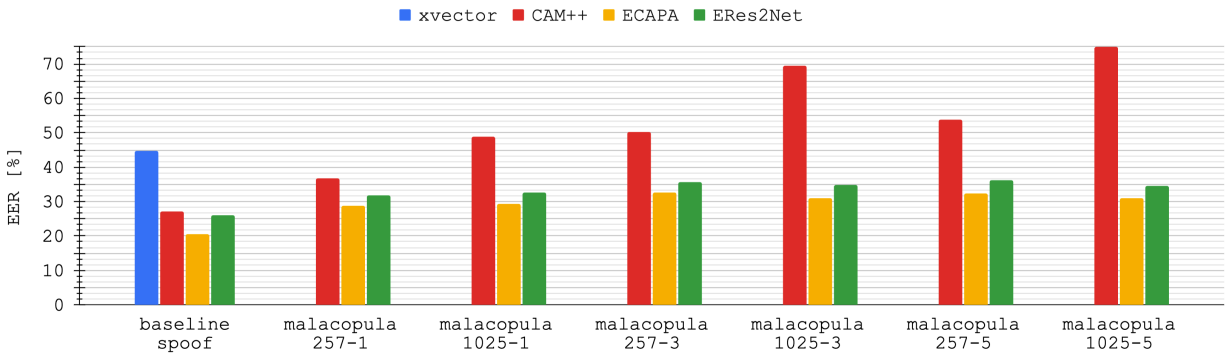

Figure 3: Pooled spf-EER for baseline spoof and Malacopula filtered spoof attacks for four different ASV systems.

dian of the distribution of spoof scores exceeds that of the target bona fide scores.

5. Experimental Setup

We use three distinct ASV systems, each with unique structural and functional characteristics. They are: CAM++ for training; ECAPA for validation; ERes2Net for testing. By employing three different ASV systems, we are able to test the transferability of Malacopula attacks across different ASV architectures and embedding extraction methodologies. Below, we provide a brief descriptions of each system.

The **CAM++** [3] system consists of a front-end convolution module and a densely connected time delay neural network (D-TDNN) backbone and the extraction of a 512-dimensional speaker embedding. Each D-TDNN layer includes a context-aware masking (CAM) module. CAM++ employs multi-granularity pooling to capture discriminative speaker characteristics, enhancing its ability to differentiate between speakers effectively.

The **ECAPA** [2] system uses the TDNN architecture, which incorporates 3 SE-Res2Block modules to extract a 192-dimensional speaker embedding. This structure leverages squeeze-and-excitation (SE) mechanisms to enhance feature representation and improve performance in speaker verification tasks.

The **ERes2Net** [4] system addresses the limitations of the traditional Res2Net architecture by integrating local and global feature fusion to extract a 192-dimensional speaker embedding. This integration allows ERes2Net to capture both detailed and holistic patterns in the input signal, enhancing its capability to recognise speaker-specific characteristics.

5.1. Database, protocols and filter optimisation

All experiments were conducted using the ASVspoof 2019 logical access (LA) dataset [25]. 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 and deepfake detectors. In contrast, the test partition contains data in the realm of the attacker. Speaker and attack-specific filters are hence trained according to (4) using the test partition, i.e. using A07 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 which is to their advantage.

Malacopula filters are trained using spoofed and bona fide utterances sourced from the *test* data partition and using CAM++ for $f_A(\cdot)$ and ECAPA as $f_B(\cdot)$, while testing is performed using ERes2Net. The setup reflects a scenario in which filters are trained by an attacker offline and then used to implement online/real-time attacks, e.g. in a logical access or telephony scenario.

5.2. Implementation

The objective function (4) is optimised with Adam [28]. Filters are optimised for 60 epochs with a batch size of 12. ASV model weights are kept frozen during optimisation. We used two filter lengths $L \in \{257, 1025\}$ and three filter depths $K \in \{1, 3, 5\}$ to explore the balance between optimisation of (4), attack success and the preservation of speech quality. Our specific implementation, along with audio examples, is available as open-source and can be used to reproduce our results under the same GPU environment.[4]

5.3. Metrics

All results are obtained using the standard SASV evaluation protocol [29] and are expressed in terms of spf-EERs computed using target (positive class) and spoofed (negative class) utterances.

6. Experimental Results

Results presented in Fig. 3 show pooled spf-EERs for the three ASV systems, comparing baseline spoof results with those using Malacopula filters of different length and depth. For comparison with a legacy ASV system, we include performance for an x-vector system among the baseline spoof results to show that modern systems are less vulnerable. For Malacopula attacks, the vulnerability increases universally, and more significantly for CAM++ which is used for training. Still, spf-EERs are higher for ECAPA and ERes2Net systems than for the baseline condition. This shows that Malacopula filters exhibit some generalisation across different ASV architectures.

Fig. 4 show a performance comparison for baseline spoof and Malacopula 257-5 filters for the three systems — CAM++, ECAPA, and ERes2Net — for all thirteen underlying spoof at-

[4] https://github.com/eurecom-asp/malacopula

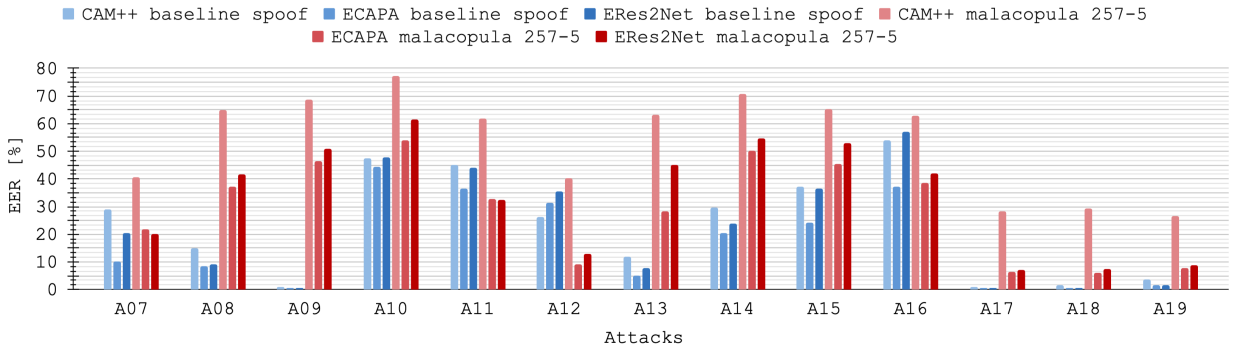

Figure 4: spf-EER per attacks for baseline spoof and Malacopula 257-5 filtered spoof attacks of three ASV systems.

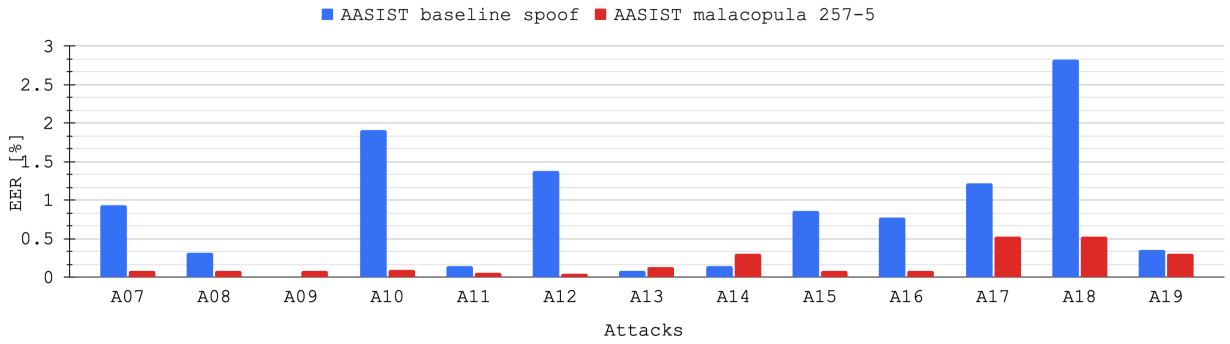

Figure 5: spf-EER per attacks for baseline spoof and Malacopula 257-5 filtered spoof attacks of three ASV systems.

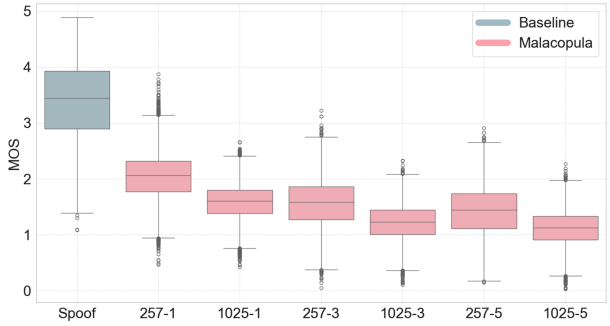

Figure 6: MOS distributions for baseline spoof and Malacopula filtered spoof.

tacks. For certain attacks, such as A09, Malacopula leads to a significant increase in the vulnerability of all three systems. However, Malacopula exhibits lower performance against certain already-effective attacks, such as A12. For attacks A17, A18, and A19, which are all voice conversion based attacks, a similar increase in vulnerability is observed.

Overall, results show that Malacopula filters provoke increased vulnerabilities across the three ASV systems and attack scenarios. This underscores the importance of continuous improvement and adaptation in ASV system defences to maintain robustness against evolving adversarial techniques.

Fig. 5 shows the impact of Malacopula 257-5 filters upon the popular ASSIST [30] spoof and deepfake detection system. Results are shown in terms of spf-EERs for baseline spoof (blue bars) and Malacopula attacks (red bars). When the utterances are processed by Malacopula, AASIST performance *improves* almost universally. Only for A13 and A14 are spf-EERs higher, albeit only very marginally, and are in any case still low. These results indicate that Malacopula attacks are easily detectable, reinforcing the need for dedicated detection solutions in order to protect ASV systems from manipulation.

Fig. 6 illustrates the impact on speech quality measured in terms of the mean opinion score (MOS) for various Malacopula configurations. All scores were estimated automatically using the method described in [31]. MOS distributions are shown for the baseline spoof and Malacopula attacks (L, K). As expected, distributions for the baseline spoof attacks are generally higher, with distribution modes of around 3 and 4, For Malacopula attacks, distribution modes are between 1 and 2. Variations in speech quality caused by Malacopula attacks are attributed to the use of more or less aggressive filters, where smaller values of L and K cause less degradation.

However, the controlled conditions under which ASVspoof 2019 source data was initially collected, do not reflect factors such as background or channel noise, which typify conditions in the wild and which may influence results. Perturbations introduced by Malacopula themselves resemble background or channel noise. This suggests the need for further investigations to verify detection performance in more realistic acoustic conditions and scenarios.

7. Conclusions

In this paper, we introduce Malacopula, an adversarial perturbation model in the form of generalised Hammerstein framework, which acts upon a speech utterance in order to exaggerate and

exploit the vulnerabilities of automatic speaker verification systems to spoofing and deepfake attacks. Malacopula extends the capabilities of previous models, enabling more effective manipulation of the amplitude, phase, and frequency components of speech signals in non-linear fashion.

Experiments, performed using the ASVspoof 2019 dataset show that Malacopula significantly increases the vulnerability of CAM++, ECAPA, and ERes2Net ASV systems to spoofing and deepfake attacks. The cross-system training and evaluation nature of the experiments underscores the robustness and transferability of Malacopula attacks, highlighting the potential threat in real-world scenarios.

Despite the power of Malacopula in increasing the threat of spoofing attacks, our analysis reveals that the resulting perturbations reduce speech quality, as reflected by lower mean opinion scores. Reassuringly, though, spoofing and deepfake detection systems like AASIST are capable of detecting Malacopula attacks. However, we acknowledge that our current work shows only that attacks are detected effectively under controlled conditions. This suggests the need for further investigations to determine whether the same defences remain robust in unconstrained scenarios. Our findings highlight the importance of continuing the hunt for new vulnerabilities and efforts to tackle them so as to ensure the reliability of ASV systems in the wild.

8. Acknowledgements

This work is supported with funding received from the French Agence Nationale de la Recherche (ANR) via the BRUEL (ANR-22-CE39-0009) and COMPROMIS (ANR-22-PECY-0011) projects.

9. References

[1] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[2] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *INTERSPEECH 2020*, 2020, pp. 3830–3834.

[3] Haibo Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen, "CAM++: A fast and efficient network for speaker verification using context-aware masking," in *INTERSPEECH 2023*, 2023.

[4] Yafeng Chen et al., "ERes2NetV2: Boosting short-duration speaker verification performance with computational efficiency," *arXiv preprint arXiv:2406.02167*, 2024.

[5] Maros Jakubec, Roman Jarina, Eva Lieskovska, and Peter Kasak, "Deep speaker embeddings for speaker verification: Review and experimental comparison," *Engineering Applications of Artificial Intelligence*, vol. 127, pp. 107232, 2024.

[6] Xuechen Liu, Md Sahidullah, Kong Aik Lee, and Tomi Kinnunen, "Generalizing speaker verification for spoof awareness in the embedding space," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1261–1273, 2024.

[7] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[8] Nicholas Carlini and David Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE Security and Privacy Workshops (SPW)*, 2018, pp. 1–7.

[9] Michele Panariello, Wanying Ge, Hemlata Tak, Massimiliano Todisco, and Nicholas Evans, "Malafide: a novel adversarial convolutive noise attack against deepfake and spoofing detection systems," in *Proc. INTERSPEECH 2023*, 2023, pp. 2868–2872.

[10] Fatima-Zahra Chaoui, Fouad Giri, Youssef Rochdi, Mohamed Haloua, and Abdessamad Naitali, "System identification based on Hammerstein model," *International Journal of Control*, vol. 78, no. 6, pp. 430–442, 2005.

[11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2015.

[12] Paarth Neekhara, Shehzeen Hussain, Prakhar Pandey, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar, "Universal adversarial perturbations for speech recognition systems," in *Proc. Interspeech 2019*, 2019, pp. 481–485.

[13] Yi Xie, Cong Shi, Zhuohang Li, Jian Liu, Yingying Chen, and Bo Yuan, "Real-time, universal, and robust adversarial attacks against speaker recognition systems," in *Proc. ICASSP 2020*, 2020, pp. 1738–1742.

[14] Xingyu Zhang, Xiongwei Zhang, Wei Liu, Xia Zou, Meng Sun, and Jian Zhao, "Waveform level adversarial example generation for joint attacks against both automatic speaker verification and spoofing countermeasures," *Engineering Applications of Artificial Intelligence*, vol. 116, pp. 105469, 2022.

[15] Andre Kassis and Urs Hengartner, "Practical attacks on voice spoofing countermeasures," *arXiv preprint arXiv:2107.14642*, 2021.

[16] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard, "Universal adversarial perturbations," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[17] Jiaqi Li, Li Wang, Liumeng Xue, Lei Wang, and Zhizheng Wu, "An initial investigation of neural replay simulator for over-the-air adversarial perturbations to automatic speaker verification," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4635–4639.

[18] Paarth Neekhara, Shehzeen Hussain, Prakhar Pandey, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar, "Universal adversarial perturbations for speech recognition systems," in *Proc. Interspeech 2019*, 2019, pp. 481–485.

[19] Weiyi Zhang, Shuning Zhao, Le Liu, Jianmin Li, Xingliang Cheng, Thomas Fang Zheng, and Xiaolin Hu, "Attack on practical speaker verification system using universal adversarial perturbations," in *Proc. ICASSP 2021*, 2021, pp. 2575–2579.

[20] Simon Grimm and Jürgen Freudenberger, "Hybrid Volterra and Hammerstein modelling of nonlinear acoustic systems," in *Fortschritte der Akustik: DAGA 2016, Aachen: 14.-17. März 2016: 42. Jahrestagung für Akustik, Tagungsband*. Dt. Gesellschaft für Akustik eV, 2016, pp. 1167–1170.

[21] K. Lashkari, "A modified Volterra-Wiener-Hammerstein model for loudspeaker precompensation," in *Conference Record of the Thirty-Ninth Asilomar Conference onSignals, Systems and Computers, 2005.*, 2005, pp. 344–348.

[22] Giovanni L. Sicuranza and Alberto Carini, "On the accuracy of generalized Hammerstein models for nonlinear active noise control," in *2006 IEEE Instrumentation and Measurement Technology Conference Proceedings*, 2006, pp. 1411–1416.

[23] Pietro Burrascano, Alessandro Terenzi, Stefania Cecchi, Matteo Ciuffetti, and Susanna Spinsante, "A swept-sine-type single measurement to estimate intermodulation distortion in a dynamic range of audio signal amplitudes," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[24] Leela K. Gudupudi, Christophe Beaugeant, and Nicholas Evans, "Characterisation and modelling of non-linear loudspeakers," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 134–138.

[25] Xin Wang et al., "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, pp. 101114, 2020.

[26] Hemlata Tak, Madhu R. Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans, "RawBoost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6382–6386.

[27] Junyan Wu, Qilin Yin, Ziqi Sheng, Wei Lu, Jiwu Huang, and Bin Li, "Audio multi-view spoofing detection framework based on audio-text-emotion correlations," *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2024.

[28] D. P. Kingma and J. Lie Ba, "Adam: A method for stochastic optimization," in *Proc. of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[29] J.-w. Jung, H. Tak, H.-j. Shim, H.-S. Heo, B.-J. Lee, S.-W. Chung, H.-J. Yu, N. Evans, and T. Kinnunen, "SASV 2022: The first spoofing-aware speaker verification challenge," in *Proc. Interspeech 2022*, 2022, pp. 2893–2897.

[30] 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," in *Proc. ICASSP 2022*, 2022, pp. 6367–6371.

[31] Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi, "Generalization ability of MOS prediction networks," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8442–8446.