LECTURE NOTES: *Deepfakes and Adversarial Threats to Automatic Speaker Verification, Countermeasures and Human Listeners*

By Massimiliano Todisco, Professor of Audio and Speech Technologies, EURECOM, France

Massimiliano Todisco, a professor of audio and speech technologies at EURECOM in France, is best known for his contributions to fake audio detection through the invention of *constant Q cepstral coefficient*. These features were widely used in speech spoofing detection before the rise of deep learning, and his work earned him the ISCA 2020 Award for the best article in *Computer Speech and Language* for the quinquennium 2015-2019. His research is dedicated to advancing voice biometrics, deepfake detection, and privacy preservation. Currently, his projects include adversarial learning strategies, and leveraging generative AI to improve both the transparency and effectiveness of audio and speech technologies. Todisco is currently co-organising two international challenges, the ASVspoof and VoicePrivacy, and he serves as an Associate Editor for the *IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM)*.

The rapid evolution of voice cloning technologies has revolutionised diverse applications and offered significant advancements, from empowering individuals with speech impairments to enhancing human-machine interactions through virtual assistants [1,2,3]. These new technologies are also transforming creative industries like gaming and dubbing. By analysing voice characteristics, these systems verify users in a manner that is both convenient and non-invasive. However, these benefits come with an alarming downside. Their misuse poses severe challenges to security systems, particularly to Automatic Speaker Verification (ASV) [4] systems that play a critical role in secure authentication for applications ranging from banking to personal devices. This reliance on voice as a biometric factor has made ASV systems a prime target for attackers. Critical vulnerabilities have been exposed through the rising number of spoofing attacks, where cloned voices mimic legitimate users, and adversarial attacks, where subtle perturbations deceive ASV systems.

What makes this problem even more pressing is that these voice cloning technologies are not limited to deceiving machines. They can also fool humans [5]. Advanced systems can now generate synthetic audio that is nearly indistinguishable from real speech [6], even to trained listeners. Moreover, voice cloning tools are widely available online, with many platforms offering accessible interfaces that require no expertise in audio engineering. This popularisation of voice cloning technology means that malicious actors, even those with minimal technical skills, can easily exploit these tools to impersonate individuals, commit fraud, or spread misinformation.

This lecture examines the current state of ASV and Countermeasure (CM) systems, focusing on the adversarial challenges posed by two recent groundbreaking attack models. By exploring their mechanisms, impacts, and implications, we aim to shed light on the vulnerabilities of ASV systems, and the pressing need for innovative secure voice-based technology solutions in an era of rapidly advancing and widely accessible voice cloning threats.

Deepfakes and Presentation Attacks: A Growing Threat to ASV

As mentioned above, ASV authentication systems face significant threats from deepfake and presentation attacks. Both attack types undermine ASV security and human trust in voice-based interactions, albeit with slightly different objectives.

Deepfakes involve the creation of highly realistic synthetic audio that mimics a target speaker's voice. Techniques like Voice Conversion (VC) [7] and Text-to-Speech (TTS) [8] synthesis enable attackers to replicate a speaker's vocal characteristics with astounding accuracy. Alarmingly, today's technologies require only a few seconds of audio from the victim to generate convincing deepfake voices.

These attacks can extend beyond ASV systems to pose a significant risk to human perception. For example, attackers can impersonate individuals in phishing scams or create speeches to spread misinformation. The increasing inability of humans to consistently distinguish between real and synthetic voices makes deepfakes a powerful tool for fraud, manipulation, and reputational harm.

Presentation attacks [9], or spoofing, target ASV systems through manipulated audio designed to bypass authentication. These attacks exploit voice cloning technologies to

mimic legitimate users and compromise ASV security. Spoofing is particularly dangerous in critical applications, such as biometric authentication for banking or personal devices, where successful attacks can lead to unauthorised access and significant harm.

The dual threat of deepfakes and presentation attacks lie in their ability to deceive both humans and machines. While deepfakes target trust in voice communication, spoofing directly undermines the reliability of ASV systems.

Adversarial Attacks: A New Frontier in ASV and CM Threats

Adversarial attacks [10] add an extra threat layer to ASV and CM systems as, unlike presentation attacks or deepfakes, they manipulate the audio signal itself. Attackers exploit specific vulnerabilities in ASV and CM algorithms by introducing subtle perturbations designed to remain imperceptible to human listeners, while significantly degrading performance.

Adversarial attacks target the decision-making processes of these systems to deceive classification or verification mechanisms. The result is to render ASV incapable of distinguishing between bona fide [(1)] and spoofed speech [9] or to cause misclassification of legitimate users. Advanced adversarial attacks are particularly effective when they adapt the noise to real-world scenarios, such as audio transmitted over communication channels where compression and noise are unavoidable.

Current Landscape of Defences: Countermeasures for Robust Speaker Verification

To address these threats, researchers have developed various countermeasures to enhance the robustness of ASV systems. Initiatives, such as the ASVspoof challenges series [11], play a critical role in promoting the development and benchmarking of these defences. By providing a standardised framework for evaluating countermeasures against diverse attack scenarios, ASVspoof fosters innovation and collaboration within the research community. Despite significant progress though, many countermeasures struggle to effectively generalise to unseen attack types or conditions. This limitation often stems from the evolving nature of voice cloning and adversarial techniques, the complexity of real-world transmission environments, and the inherent variability in human speech.

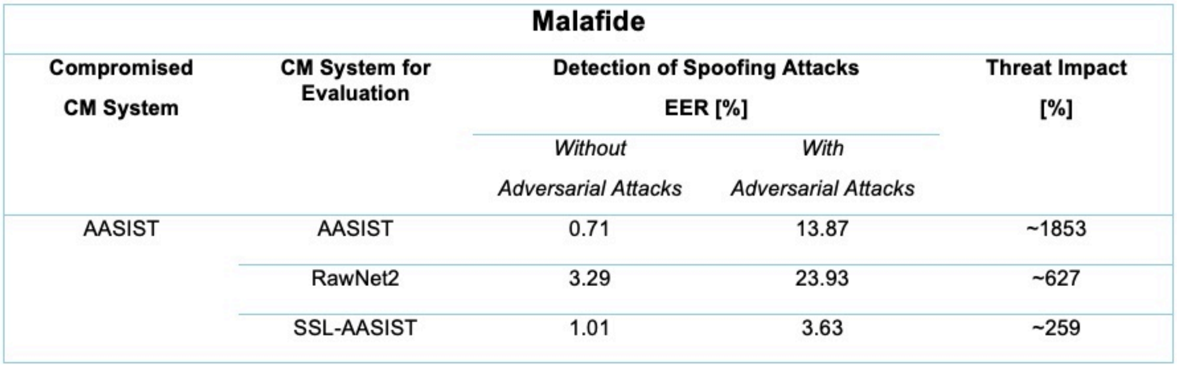
Advanced Adversarial Techniques: Malafide and Malacopula [(2)]

Malafide [12] and Malacopula [13] are two groundbreaking adversarial techniques that exploit weaknesses in existing defense systems through tailored perturbations. This makes them particularly robust in practical conditions. Malafide demonstrates a significant threat to anti-spoofing countermeasure (CM) systems used to secure voice biometrics. By employing a linear time-invariant filter, this strategy introduces convolutive noise that deceives systems into misclassifying spoofed speech as bona fide. Unlike traditional methods that rely on utterance-specific noise, Malafide is adapted to specific attack scenarios, enabling real-time application and revealing critical vulnerabilities in widely used CM systems. Its key features include:

- **Universal Applicability**: Malafide's versatile filter is pre-trained and operates independently of the speech's duration or content.
- **Efficient Deployment:** It is computationally lightweight and acts as a post-processing filter, making it applicable for real-time application.
- **Cross-System Transferability**: The generated perturbations generalise across different CM architectures and utterances.
- **Human Perception:** Added perturbations resemble conventional audio equalisation or reverberation effects, making it difficult to detect them as malicious.

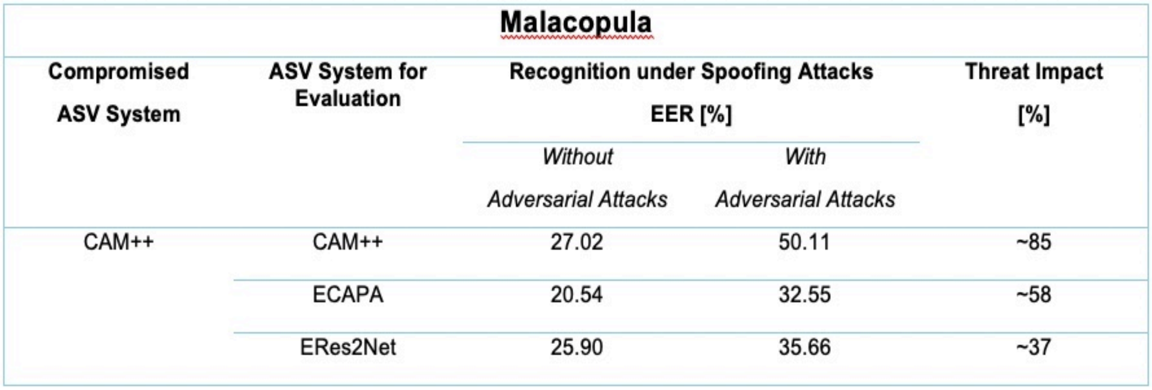
Building on Malafide, Malacopula employs a generalised Hammerstein model [14], combining non-linear transformations with convolutive filtering. This approach allows for more complex manipulation of the audio signal, targeting amplitude, phase, and frequency components to create perturbations that are both highly effective and difficult to detect. Malacopula is specifically tailored for speaker- and attack-specific scenarios, optimising its perturbations to bring spoofed speech and the target speaker closer together. This capability makes it uniquely effective at deceiving ASV systems under spoofing attacks, even when advanced countermeasures are employed. Like Malafide, Malacopula excels in codec-based communication channels by introducing perturbations that remain resilient to compression artefacts. Its key features include:

- **Attack Optimization**: Malacopula minimises the cosine distance between the embeddings of processed spoofed speech with adversarial noise and bona fide speech, ensuring precise deception.
- **Cross-System Transferability**: The perturbations generalise across different ASV architectures and utterances.



Malafide				
Compromised CM System	CM System for Evaluation	Detection of Spoofing Attacks EER [%]		Threat Impact [%]
		Without Adversarial Attacks	*With Adversarial Attacks*	
AASIST	AASIST	0.71	13.87	~1853
	RawNet2	3.29	23.93	~627
	SSL-AASIST	1.01	3.63	~259

Table 1: CM systems experience dramatic EER increases, with threats rising up to 1853%. This demonstrates Malafide's ability to render CMs highly ineffective.



Malacopula				
Compromised ASV System	ASV System for Evaluation	Recognition under Spoofing Attacks EER [%]		Threat Impact [%]
		Without Adversarial Attacks	*With Adversarial Attacks*	
CAM++	CAM++	27.02	50.11	~85
	ECAPA	20.54	32.55	~58
	ERes2Net	25.90	35.66	~37

Table 2: ASV systems show EER increases up to 85%, highlighting Malacopula's precision in embedding manipulation.

- **Real-World Effectiveness***:* Malacopula's perturbations remain effective after transmission and compression, making it highly practical for telephony use cases.
- **Lightweight Design***:* Despite its complexity, Malacopula is efficient and deployable in real-time scenarios.

These innovations uniquely position Malafide and Malacopula to challenge ASV and CM systems, particularly in practical scenarios involving codec-compressed audio. The perturbations they introduce resemble artefacts caused by compression and transmission, ensuring the attack remains effective even after codec processing. This property makes Malafide highly practical and dangerous for telephony and VoIP scenarios.

Impact of Malafide and Malacopula on ASV and CM Systlems

Results from evaluations on the ASVspoof 2019 LA database [15] demonstrate the significant impact of Malafide and Malacopula on ASV and CM systems. Both methods have proven their ability to significantly degrade the performance of even state-of-the-art ASV, such as CAM++

[16], ECAPA [17], and ERes2Net [18], and CM architectures, such as RawNet2 [19], AASIST [20] and SSL-AASIST [21]. Tables 1 and 2 above summarise the impact of Malafide and Malacopula on ASV and CM systems.

Conclusions and Future Directions

The results of Malafide and Malacopula as documented above highlight critical vulnerabilities in ASV and CM systems. The evolving landscape of adversarial attacks demands proactive and innovative approaches to ensure the security of ASV and CM systems. As attacks like Malafide and Malacopula demonstrate, the combination of codec resilience, lightweight design, and practical deployment poses a significant challenge to current defences.

These factors highlight the need for adaptive solutions capable of addressing a wide range of attack scenarios. The integration of adversarial training, neural codec-specific defences, and self-supervised learning methods offers promising avenues for overcoming these obstacles. Yet, the rapid evolution of attack methods underscores that the security of ASV and CM systems remains an ongoing challenge, one that requires continuous innovation. We must bear in mind that the dual goal is to ensure that these systems are not only resilient to advanced attacks, but also maintain their reliability and usability in diverse real-world applications.

Endnotes

1) *Bona fide* is Latin for "good faith." It signifies sincerity, authenticity, or genuine intention without deceit or fraud.
2) *Mala fide* is Latin for “in bad faith.” It signifies actions or intentions that are deceitful, dishonest, or intended to mislead or harm. *Mala copula* is Latin for “bad connection” or “bad union.” It signifies an undesirable or improper association between elements.

References

1) S.C. Ramu, D. Saxena and V. Mali, "A Survey on Voice Cloning and Automated Video Dubbing Systems," *International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, Chennai, India, 2024, pp. 1-5. doi: 10.1109/WiSPNET61464.2024.10532876.
2) M.A.M. Ahmed, K.A. Elghamrawy and Z.A. El Haliem Taha, "(Voick): Enhancing Accessibility in Audiobooks Through Voice Cloning Technology," *6th International Conference on Computing and Informatics (ICCI)*, New Cairo - Cairo, Egypt, 2024, pp. 46-52, doi: 10.1109/ICCI61671.2024.10485044.
3) A. Pérez, G.G. Díaz-Munío, A. Giménez, J.A. Silvestre-Cerdà, A. Sanchis, J. Civera, M. Jiménez, C. Turró, and A. Juan, “Towards Cross-lingual Voice Cloning in Higher Education,” *Engineering Applications of Artificial Intelligence*, Volume 105, 2021, 104413, ISSN 0952-1976.

4) M. Jakubec, R. Jarina, E. Lieskovska, and P. Kasak, “Deep Speaker Embeddings for Speaker Verification: Review and Experimental Comparison,” *Engineering Applications of Artificial Intelligence*, Volume 127, Part A 2024, 107232, ISSN 0952-1976.

5) M. Krzysztof, S. Zaporowski, and A. Czyżewski. "Comparison of the Ability off Neural Network Model and Humans to Detect a Cloned Voice." *Electronics* 12.21 (2023): 4458.

6) H-W. Yoon et al., "Enhancing Multilingual TTS with Voice Conversion Based Data Augmentation and Posterior Embedding," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Seoul, Korea 2024, pp. 12186-12190, doi: 10.1109/ICASSP48485.2024.10448471.

7) N. Guo, J. Wei, Y. Li, W. Lu, J. Tao, “Zero-shot voice conversion based on feature disentanglement,” *Speech Communication*, Volume 165, 2024, 103143, ISSN 0167-6393.

8) X. Tan et al., "NaturalSpeech: End-to-End Text-to-Speech Synthesis With Human-Level Quality," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4234-4245, June 2024, doi: 10.1109/TPAMI.2024.3356232.

9) *ISO/IEC 30107-1:2023. Information Technology — Biometric Presentation Attack Detection, Part 1: Framework*, Published, Edition 2, 2023.

10) A. Kurakin, et al. "Adversarial Attacks and Defences Competition," *The NIPS '17 Competition: Building Intelligent Systems*. Springer International Publishing, 2018.

11) X. Wang, H. Delgado, H. Tak, J. Jung, H. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen, N. Evans, K.A. Lee, and J. Yamagishi, “ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale,” *ASVspoof Workshop 2024*, 31 August 2024, Kos, Greece.

12) M. Panariello, W. Ge, H. Tak, M. Todisco and N. Evans, “Malafide: A Novel Adversarial Convolutive Noise Attack against deepfake and spoofing Detection Systems,” *INTERSPEECH 2023*, 20-24 August 2023, Dublin, Ireland.

13) M. Todisco, M. Panariello, X. Wang, H. Delgado, K.A. Lee, and N. Evans, “Malacopula: Adversarial automatic Speaker Verification Attacks using a Neural-based Generalised Hammerstein Model,” *ASVspoof Workshop 2024*, 31 August 2024, Kos, Greece.

14) S. Grimm and J. Freudenberger, “Hybrid Volterra and Hammerstein Modelling of Nonlinear Acoustic Systems,” in *Fortschritte der Akustik: DAGA 2016.*

15) M. Todisco, X. Wang et al., “ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection,” in *Proceedings of Interspeech,* 2019.

16) H. Tak, J. Patino et al., “End-to-end Anti-spoofing with RawNet2,” in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP*) 2021.

17) J-w. Jung, H.S. Heo et al., “AASIST: Audio Anti-spoofing using Integrated Spectro-temporal Graph Attention

Networks," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP*), 2022.

18) H. Tak, M. Todisco et al., "Automatic Speaker Verification Spoofing and Deepfake Detection using wav2vec 2.0 and Data Augmentation," in *Proceedings of the. Speaker Odyssey Workshop,* 2022.

19) H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "CAM++: A Fast and Efficient Network for Speaker Verification using Context-aware Masking," in *Proceedings of INTERSPEECH 2023*, 2023.

20) B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN based Speaker Verification," in *Proceeding of INTERSPEECH 2020.*

21) Y. Chen et al., "ERes2NetV2: Boosting short-duration speaker verification performance with computational efficiency," in *Proceedings of INTERSPEECH 2024.*

NOTED IN THE LITERATURE

Spatio-Temporal Dual-Attention Transformer for Time-Series Behavioral Biometrics

A summary of an article that appeared in IEEE Transactions on Biometrics, Behavior, and Identity Science in October, 2024, as prepared by its authors Kim-Ngan Nguyen, Sanka Rasnayaka, Sandareka Wickramanayake, Dulani Meedeniya, Sanjay Saha, and Terence Sim

INTRODUCTION

Recent advancements in mobile technology have enabled people to easily carry out crucial tasks, such as communication, finance, and healthcare, via their smartphones. With these advances come the need for secure, yet user-friendly, authentication methods. One-time/session-based authentication systems—including knowledge-based authentication methods that utilize pin codes or passwords, or physiological biometrics using fingerprints or faces—require user involvement that can reduce the ease of use. Continuous Authentication (CA) offers a solution by verifying users based on their behavioral patterns, like keystrokes and swipes, as they use a device. Additionally, the integration of affordable IMU sensors in most smartphones today also enhances CA by providing additional data for more accurate