IberSPEECH 2024
11-13 November 2024, Aveiro, Portugal

Signal and Neural Processing against Spoofing Attacks and Deepfakes for Secure Voice Interaction (ASASVI)

Angel M. Gomez[1], *Antonio M. Peinado*[1], *Victoria E. Sánchez*[1], *Iván López-Espejo*[1], *Alejandro Gómez-Alanis*[1], *Eros Roselló*[1], *José C. Sánchez-Valera*[1], *Juan M. Martín-Doñas*[2]

[1]Dept. Signal Processing, Telematics and Communications - CITIC, Universidad de Granada, Spain
[2]Fundación Vicomtech, Basque Research and Technology Alliance (BRTA),
Mikeletegi 57, 20009 Donostia-San Sebastián (Spain)

{amgg,amp,victoria,iloes,agomezalanis,erosrosello,svjosecarlos}@ugr.es,
jmmartin@vicomtech.org

Abstract

The increasing sophistication of multimodal interaction systems, which enable human-like communication, raises concerns regarding the authenticity of exchanged speech data. Our research addresses the challenges posed by the malicious misuse of speech technologies, as for example voice conversion (VC) and text-to-speech (TTS), which can be exploited to impersonate speakers, manipulate public opinion, or compromise voice biometric systems. Existing countermeasures, known as anti-spoofing techniques, face significant limitations in effectively combating these threats. To tackle this, our project proposes three research directions: (1) improving deep neural network (DNN)-based anti-spoofing techniques through robust feature extractors, novel architectures, and enhanced training methodologies to bridge the gap between laboratory performance and real-world application, (2) generating more realistic and diverse training data to better reflect real-world conditions and attacks, and (3) developing advanced, imperceptible watermarking techniques for synthesized speech to prevent misuse, even in the presence of deep learning-based removal attempts. This research aims to significantly enhance the security and reliability of computer-mediated speech interactions.

Index Terms: Speech-based interaction, Security, Voice biometrics systems, Voice impersonation, Anti-spoofing, Deepfakes, Artificial intelligence

1. Introduction

The rapid advancement of digital technologies has ushered in a new era of human-computer interaction, allowing us to engage with information systems in a multimodal manner, where systems can communicate with us as if they were human. While this enhances user experience and accessibility, it also introduces significant challenges regarding the authenticity of the exchanged data. Thus, crucial questions arise: Is the data received authentic? Can we trust its origin? This issue becomes particularly concerning when the primary medium of interaction is speech, as advanced technologies now enable the creation of highly realistic and natural-sounding voices, which can be manipulated for malicious purposes.

Voice-based interactions are now vulnerable to sophisticated methods such as voice conversion (VC) and text-to-speech (TTS) systems. These technologies, while groundbreaking, can be misused to impersonate someone's voice, leading to security breaches and fraud. VC allows the transformation of a speech recording from one individual to another, while TTS can generate speech mimicking a person's voice based on a small reference sample —often requiring as little as three seconds of audio [1]. Although these technologies have legitimate applications in assistive technologies, virtual assistants, and entertainment industries [2, 3], they also pose a significant risk of being misused. Misrepresentation, manipulation, and misinformation through fake voice recordings have already been identified as serious threats, along with more direct forms of fraud, such as impersonating someone in a phone call to deceive others [4, 5].

One domain that is particularly susceptible to these attacks is voice biometric systems, which are increasingly used for authentication purposes. Voice authentication offers a natural and convenient method of identity verification, but it can be easily compromised. Despite the maturity of automatic speaker verification (ASV) technology [6], it remains vulnerable to a wide range of spoofing attacks. Techniques such as mimicking, replaying speech recordings, or using advanced speech synthesis and conversion can effectively bypass these systems.

Spoofing attacks are typically classified into physical access (PA) and logical access (LA) categories [7]. PA attacks involve replaying recorded speech using the system's microphone, while LA attacks involve synthetic or converted voices directly injected into the system. Audio deepfake (DF) detection is inherently tied to LA attack detection, as both rely on identifying high-quality forged speech from TTS or VC systems. The main difference lies in the target, either deceiving a human listener or an ASV system. Since 2015, a number of initiatives and challenges [8–10] have driven the development of countermeasures, also known as anti-spoofing techniques, to detect and mitigate these impersonation attacks [11]. However, despite the considerable progress made, significant challenges remain.

Deep neural network (DNN) based systems have become essential in detecting spoofing attacks, surpassing traditional signal processing methods [12–14]. These systems typically use neural networks to extract deep features and generate an identity vector or embedding for the entire audio sequence. Then, a binary classifier determines whether the input is genuine or spoofed. In practice, these components are often integrated into a single network (see Figure 1), making them more efficient but still requiring improvements to detect increasingly sophisticated attacks.

Despite DNNs have shown great potential in modeling complex relationships within speech data, their real-world performance often falls short. One key issue is that DNNs tend to overfit to training data, especially when only limited or non-

10.21437/IberSPEECH.2024-49

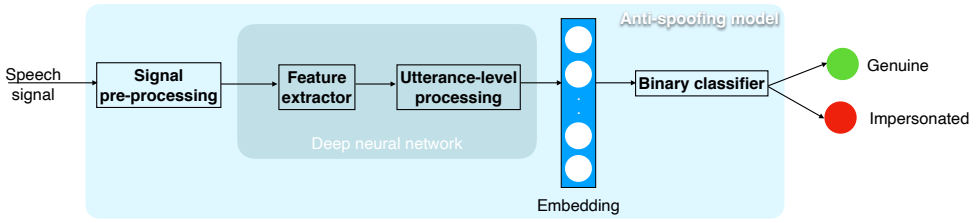

Figure 1: *Diagram of a typical anti-spoofing system. The input speech undergoes pre-processing before being passed to a deep feature extractor, which generates an embedding representing the entire utterance. A binary classifier then determines whether the speech is genuine or spoofed. In many cases, these components are integrated into a single DNN architecture.*

representative datasets are available. As spoofing attacks can vary widely, making it nearly impossible to account for all possible attack vectors in training data, DNN-based anti-spoofing methods often struggle to generalize to real-world scenarios. This highlights the need for more robust or, even, alternative solutions.

2. Project goals

This project focuses on enhancing the security of voice-based interactions, particularly in the contexts of voice biometrics and audio deepfakes. Building upon our previous research [15], this project aims to address current anti-spoofing systems shortcomings. To this end, we will first explore novel DNN architectures, cost functions and training strategies. Explainability will be also crucial for understanding how models detect genuine or spoofed speech, providing insight into the decision-making process. Additionally, the environmental impact of DNN models will be considered, as their computational requirements have skyrocketed, leading to significant increases in energy consumption and carbon emissions [16].

Another critical aspect of this project is the creation of improved anti-spoofing databases. High-quality, diverse datasets are essential for training models that can generalize to a wide range of attack scenarios. The evolution of anti-spoofing challenges has shown that generating more realistic data that reflect real-world conditions is crucial, as is employing data augmentation techniques to enhance diversity within training sets.

Finally, as deep learning methods for generating synthetic speech continue to evolve, there is a risk that detection systems will become less effective against increasingly realistic deepfakes. In anticipation of this, we propose investigating robust and imperceptible watermarking techniques for synthesized speech, ensuring that deepfake audio can be reliably tagged and identified, even if attackers attempt to tamper with the watermark. This will help create a more secure framework for speech-based interactions in the digital age.

2.1. Feature extractors and detection models

Speech features play a critical role in detecting spoofing attacks in voice biometrics systems. Traditional speech features such as filter banks (FBANK) and Mel-frequency cepstral coefficients (MFCC) have been replaced by more advanced extractors designed specifically for spoof detection, such as constant-Q cepstral coefficients (CQCC) [17] and long-term spectral statistics (LTSS) [18]. These features primarily focus on spectral magnitude, but recent approaches also explore working directly with audio samples through task-oriented deep learning models, such as convolutional neural networks (CNNs), SincNet [19], and self-supervised models like wav2vec 2.0 [20].

Recurrent networks are frequently used for extracting identity vectors and making final decisions in spoofing detection. For example, the Gated Recurrent Convolutional Neural Network [12] and RawNet2 [21] have shown strong performance against both PA and LA attacks. Emerging models now incorporate attention mechanisms and graph neural networks [22], while our team is investigating novel architectures such as Conformer [13].

Beyond architecture, the choice of a loss function is crucial for optimizing detection models. While cross-entropy is the standard, alternative approaches like triplet loss [23] and kernel-based methods [24] are being explored to enhance feature discrimination and generalization.

Despite these advances, current state-of-the-art systems still show significant limitations, with high equal error rates (EER) in cross-database evaluations [25], making those unsuitable for real-world applications. This project aims to address these gaps by developing more robust feature extractors, novel DNN architectures, and better loss functions. Additionally, we will prioritize energy efficiency and carbon footprint considerations (by means of tools as the proposed in [26]) and explore explainable AI (xAI) frameworks [27] to enhance both model transparency and performance.

2.2. Data generation and augmentation

One of the main challenges in developing anti-spoofing systems is the availability of suitable data, especially for PA attacks. The ASVspoof challenge series [8] has played a significant role in providing standardized datasets for training and evaluating spoof detection methods. However, earlier datasets, such as ASVspoof 2017 [7], revealed that models often exploited defects in the data generation process, leading to overfitting rather than solving the actual problem. In response, ASVspoof 2019 [28] used simulated PA data to control conditions. As a result, models achieved strong performance in the simulated environment. However, when tested on real-world data, they performed poorly, showing an inability to generalize. The ASVspoof 2021 [11] challenge continued this line, further confirming the difficulty models face in adapting to real-world spoofing scenarios, with EERs exceeding 24%. Logical access attack detection faces similar challenges in terms of generalization. While fusing complementary subsystems has shown some success in combating diverse speech synthesis attacks [28], new issues have arisen due to factors such as speech transmission effects and the inclusion of deepfake detection tasks in ASVspoof 2021.

This underscores the need for more suitable data to train robust anti-spoofing systems. However, collecting extensive real-world data can be costly, prompting the exploration of data aug-

mentation techniques. Methods such as speech companding, encoding or RawBoost [29], as well as generative adversarial networks (GANs) [30, 31] offer promising ways to extend and diversify existing datasets.

Given these challenges, this project will focus on two key areas: realistic PA data generation and LA data augmentation. Our goal is to develop data that mirrors the variability of real-world conditions while addressing the causes of overfitting and poor generalization in deep learning models. By tackling these issues in tandem, we aim to enhance the robustness and effectiveness of anti-spoofing systems.

2.3. Robust watermarking for legitimate applications

In scenarios where detection systems fail to identify voice impersonation, legitimate voice generation systems must collaborate to prevent misuse [32]. Audio watermarking offers a promising solution by imperceptibly embedding additional data within synthetic audio, enabling the tracking and identification of its origin. Even if attackers can bypass passive detection methods, the watermark remains in the audio file.

Initially developed for the music industry, audio watermarking has expanded to applications such as digital rights management and forensics, with methods offering varying levels of robustness, imperceptibility, and embedding capacity. However, while deep learning is being applied to enhance watermarking and steganography techniques [33, 34], it also poses a threat, as neural networks can remove watermarks with growing success [35]. This highlights the need for specifically designed watermarking methods that can resist such attacks.

This project will focus on developing specialized speech watermarking techniques for voice generation systems. These methods will aim to resist deep learning-based removal attempts, ensuring the synthetic nature of the speech can be verified. We will also explore how these techniques can be integrated into voice systems in a standardized way to improve security and accountability.

3. Materials and methods

We will follow a traditional methodology based on the experimental evaluation of our anti-spoofing proposals. Our hypotheses will be implemented and tested against state-of-the-art methods through computer simulations. To ensure a fair comparison, we will adopt the experimental framework from the ASVspoof challenges, which provide standardized databases for training, validation, and testing, along with baseline systems and performance metrics. The following datasets will be included in our framework:

- **ASVspoof 2017** [7]: Based on the *RedDots* corpus, this dataset includes only replay attacks and comprises training, validation, and test subsets with 1508, 950, and 12008 spoofed and 1508, 760, and 1298 genuine utterances across different speakers and recording sessions.
- **ASVspoof 2019** [28]: Features both LA and PA scenarios with 107 speakers, consisting of conversion, synthesis, and replay attacks. The evaluation includes unknown attacks for LA and PA scenarios, with 12483 and 108978 spoofed, and 28890 and 163740 genuine utterances, respectively. A small set of real replays is also included.
- **ASVspoof 2021** [11]: Focuses on new evaluation data for LA, PA, and Deepfake (DF) tasks, with no new training or development data provided. It involves 48 speakers and contains large datasets of genuine and spoofed utterances, simulating real spoof conditions.

Additionally, we will include the databases from our transference agreements with Veridas [36], the 2022 ADD challenge [10], VoxCeleb2, VCTK, and ASVspoof V [8] (once available).

Performance will be measured using EER [7] and the tandem detection cost function (t-DCF) [28]. For watermark robustness, we will use bit error rate (BER) and generalized likelihood ratio tests [37], as well as PESQ [38], ESTOI, MOS, and MUSHRA tests [39] for quality and intelligibility assessment.

Our methods will be developed in Python, preferred for its open-source nature and support for machine learning libraries like TensorFlow and PyTorch, ensuring optimized GPU computation.

4. Expected impact

Voice biometrics adoption is growing rapidly, driven by the demand for secure access methods and the proliferation of devices that simplify authentication. Organizations in fraud-prone sectors such as banking, insurance, and healthcare are particularly drawn to these systems for their enhanced security. The global voice biometrics market is projected to grow from $1.3 billion in 2021 to $4.8 billion by 2028, reflecting a compound annual growth rate of 20.6% [40]. Additionally, biometric technology used in smartphone transactions is expected to rise significantly, with contactless technologies outpacing traditional fingerprint-based methods. In Spain, companies like BBVA, Telefónica, and Banco Sabadell are already integrating voice biometrics into their security systems, with some offering advanced features such as voice signatures [41] and proof-of-life verification through voice [36].

Despite their benefits, voice biometrics systems are vulnerable to spoofing attacks. Indeed, according to Pindrop Security, up to 1 in 40 calls could be high-risk [42]. As attackers find ways to deceive voice-based authentication, reliable anti-spoofing technologies become critical for ensuring the safety of these systems. In addition to conventional spoofing, the rise of audio deepfakes poses a significant social threat. While less publicized than video deepfakes, audio deepfakes could be even more damaging due to the prevalence of verbal communication in daily life [43]. Attackers can exploit these fakes through phone calls, radio broadcasts, and voice messages, emphasizing the need for advanced methods to verify the authenticity of audio samples.

As this project aims to address these security gaps by developing robust techniques, it has significant potential for technology transfer. Indeed, it is supported by national companies like Biometric Vox and Veridas, which have expressed interest in its outcomes while there is an ongoing work with Veridas [36], underlining the project's potential for real-world impact.

5. Funding

The ASASVI project (PID2022-138711OB-I00) is funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

6. References

[1] C. Wang *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," 2023. [Online]. Available: https://arxiv.org/abs/2301.02111

[2] J. A. Gonzalez *et al.*, "Direct speech reconstruction from articulatory sensor data by machine learning," *IEEE/ACM Transactions*

on Audio, Speech, and Language Processing, vol. 25, no. 12, pp. 2362–2374, 2017.

[3] J. A. Gonzalez-Lopez *et al.*, "Silent speech interfaces for speech restoration: A review," *IEEE Access*, vol. 8, pp. 177 995–178 021, 2020.

[4] Avast, "Voice fraud scams company out of 243,000 dollars," 2024. [Online]. Available: https://blog.avast.com/deepfake-voice-fraud-causes-243k-scam

[5] Forbes, "Fraudsters cloned company director's voice in $35 million bank heist, police find," 2024. [Online]. Available: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions

[6] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[7] Z. Wu *et al.*, "ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[8] "ASVspoof: Automatic Speaker Verification and Spoofing Countermeasures Challenge," 2015, accessed on July 26th, 2024. [Online]. Available: https://www.asvspoof.org/

[9] B. Dolhansky *et al.*, "The deepfake detection challenge dataset," 2020.

[10] J. Yi *et al.*, "ADD 2022: the first audio deep synthesis detection challenge," in *2022 IEEE ICASSP*. IEEE, 2022.

[11] X. Liu *et al.*, "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.

[12] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A gated recurrent convolutional neural network for robust spoofing detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 1985–1999, 2019.

[13] E. Rosello, A. Gomez-Alanis, A. M. Gomez, and A. Peinado, "A conformer-based classifier for variable-length utterance processing in anti-spoofing," in *Proc. INTERSPEECH 2023*, 2023, pp. 5281–5285.

[14] S. H. Mun *et al.*, "Towards single integrated spoofing-aware speaker verification embeddings," in *Proc. INTERSPEECH 2023*, 2023, pp. 3989–3993.

[15] A. M. Gómez *et al.*, "Fusion of classical digital signal processing and deep learning methods (FTCAPPS)," in *IberSPEECH 2022*, 2022, pp. 237–240.

[16] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Commun. ACM*, vol. 63, no. 12, p. 54–63, Nov. 2020.

[17] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *The Speaker and Language Recognition Workshop (Odyssey 2016)*, 2016, pp. 283–290.

[18] H. Muckenhirn, P. Korshunov, M. Magimai-Doss, and S. Marcel, "Long-term spectral statistics for voice presentation attack detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2098–2111, 2017.

[19] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[20] H. Tak *et al.*, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," 2022. [Online]. Available: https://arxiv.org/abs/2202.12233

[21] ——, "End-to-end anti-spoofing with RawNet2," in *ICASSP 2021*, 2021, pp. 6369–6373.

[22] ——, "Graph attention networks for anti-spoofing," in *Interspeech 2021*, 2021, pp. 2356–2360.

[23] S. Novoselov *et al.*, "Triplet loss based cosine similarity metric learning for text-independent speaker recognition," in *Interspeech 2018*, 2018, pp. 2242–2246.

[24] A. Gomez-Alanis, J. A. Gonzalez-Lopez, and A. M. Peinado, "A kernel density estimation based loss function and its application to ASV-spoofing detection," *IEEE Access*, vol. 8, pp. 108 530–108 543, 2020.

[25] N. M. Müller *et al.*, "Does audio deepfake detection generalize?" 2024. [Online]. Available: https://arxiv.org/abs/2203.16263

[26] L. F. W. Anthony, B. Kanding, and R. Selvan, "Carbontracker: Tracking and predicting the carbon footprint of training deep learning models," ICML Workshop on Challenges in Deploying and Monitoring Machine Learning Systems, July 2020, arXiv:2007.03051.

[27] W. Ge, J. Patino, M. Todisco, and N. Evans, "Explaining deep learning models for spoofing and deepfake detection with shapley additive explanations," in *ICASSP 2022*, 2022, pp. 6387–6391.

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

[29] H. Tak *et al.*, "RawBoost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *ICASSP 2022*, 2022, pp. 6382–6386.

[30] A. Gomez-Alanis, J. A. Gonzalez, and A. M. Peinado, "Adversarial transformation of spoofing attacks for voice biometrics," in *IberSPEECH 2021*, 2021, pp. 255–259.

[31] A. Gomez-Alanis, J. A. Gonzalez-Lopez, and A. M. Peinado, "GANBA: Generative adversarial network for biometric anti-spoofing," *Applied Sciences*, vol. 12, no. 3, 2022.

[32] E. Parliament, "European artifical intelligence act," 2024. [Online]. Available: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf

[33] G. Chen *et al.*, "WavMark: Watermarking for audio generation," 2024. [Online]. Available: https://arxiv.org/abs/2308.12770

[34] F. Kreuk *et al.*, "Hide and speak: Towards deep neural networks for speech steganography," 2020. [Online]. Available: https://arxiv.org/abs/1902.03083

[35] E. Quiring, D. Arp, and K. Rieck, "Forgotten siblings: Unifying attacks on machine learning and digital watermarking," in *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2018, pp. 488–502.

[36] A. Peinado, A. Gomez, and M. Chica, "Validación de algoritmos de antispoofing para biometría de voz en una base de datos desarrollada por la empresa bajo licencia Mozilla Public License 2.0." Contrato OTRI con Veridas Digital Authentication Solutions S.L., 2022.

[37] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," *Signal Processing*, vol. 66, no. 3, pp. 337–355, 1998.

[38] J. M. Martin-Doñas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1680–1684, 2018.

[39] P. C. Loizou, *Speech Quality Assessment*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 623–654.

[40] Research and Markets, "Global voice biometrics market analysis/forecast report 2022-2028," 2022. [Online]. Available: https://www.businesswire.com/news/home/20220826005378/en/Global-Voice-Biometrics-Market-AnalysisForecast-Report

[41] B. V. FirVox, "Validez legal del contrato por voz," 2022. [Online]. Available: https://biometricvox.com/firvox-firma-biometrica-voz/

[42] I. Pindrop Security, "2022 voice intelligence and security report," 2022. [Online]. Available: https://go.pindrop.com/resources/report/2022-voice-intelligence-and-security-report/

[43] L. Blue and P. Traynor, "Deepfake audio has a tell researchers use fluid dynamics to spot artificial imposter voices," The Conversation, 2022.