Call Verify Secure Protocol (CVSP): Network-Side Acoustic Authentication for Real-Time Voice Integrity and Anti-Spoofing

Gilberto A. Pérez Zorrilla

Professor, Universidad Dominicana O&M (O&M University), Dominican Republic
Cybersecurity Researcher (Cyberacoustics)

In the context of this work, the term *cyberacoustics* is introduced as an emerging discipline within cybersecurity, defined as:

"Cyberacoustics is the discipline that studies, analyzes and develops methods of protection, authentication, monitoring and attack related to acoustic signals in digital environments and cybernetic systems. It integrates principles of forensic acoustics, voice biometrics, signal processing, artificial intelligence and information security."

This conceptual framework allows voice to be addressed not only as a means of communication, but also as an attack surface and a verifiable identity vector within digital infrastructures.

I. Summary

This article introduces the ***Call Verify Secure Protocol (CVSP),*** a real-time voice authentication architecture implemented at the network level, based on highly discriminated acoustic representations. Unlike traditional Automatic Speaker Verification (ASV) systems, which rely on high-dimensional embeddings and full audio processing, CVSP minimizes data exposure by transforming the voice signal into compact acoustic vectors designed to be non-reversible (~1.1 KB), in alignment with Privacy by Design principles.

The protocol addresses vulnerabilities against Logical Access (LA) attacks, such as neural synthesis (TTS) and speech conversion (VC), as well as Physical Access (PA) attacks, including replay attacks. Authentication is performed using a cosine similarity metric in a standardized feature space, complemented by a cryptographic integrity layer that links the acoustic representation with device identifiers (IMEI/UUID) and time stamps using SHA-256 hash functions.

The simulation results, inspired by ASVspoof 2019 and 2021 evaluation scenarios, suggest that CVSP can achieve effective separation between genuine and falsified signals (cosine similarity $\mu \approx 0.92$ vs. $\mu \approx 0.29$–0.31), with estimated t-DCF values of 0.061 for neural synthesis attacks. In addition, the hardware-bound integrity layer significantly reduces the effectiveness of replay attacks (≈99.6% under controlled conditions).

The protocol is designed to operate transparently within telecommunications infrastructure, using existing signaling channels (NITZ/NAS) to provide indicators of trust to the user, while explicitly recognizing associated threat vectors such as fake base stations, signaling tampering, and the potential performance degradation under codec variability.

CVSP is positioned as a complementary security layer for modern telecommunications systems, integrating acoustic biometrics with trust mechanisms at the network level. Future work is aimed at its implementation in real environments and its large-scale validation.

II. Introduction

The integrity of voice communications is currently compromised by the development of generative models capable of producing high-fidelity speech synthesis (*Deepfakes*). The challenges posed by ASVspoof initiatives [1] have shown that traditional biometric systems are susceptible to spoofing attacks that mimic the spectral characteristics of the legitimate speaker.

There are two critical limitations to the current state of the art:

Latency and Computational Load: Most countermeasures (CMs) require the processing of dense audio signals, making it difficult to implement in real-time on standard telecommunications networks.

Biometric Data Privacy: The storage of raw voice templates represents a security risk and is difficult to comply with data protection regulations such as GDPR or similar frameworks.

The CVSP protocol addresses these limitations through a workflow that transforms the acoustic signal into a vector of high-discrimination characteristics without reversibility to audio. The main contributions of this work are:

- A Privacy by Design model that prevents the reconstruction of the acoustic identity from the generated certificate.
- A geometric validation method based on Cosine Similarity applied to high-resolution acoustic parameters.
- A hardware security scheme that prevents replay attacks by digital signature of metadata from the sending device.
- An explicit discussion of the residual threat model associated with the NITZ/NAS signaling mechanism.

From the perspective of cyberacoustics, this approach redefines voice as a verifiable element within network infrastructures, where the physical properties of the acoustic signal are used as the basis for authentication and security mechanisms.

III. Theoretical Framework and Related Works

The distinction between genuine acoustic signals and spoofing attacks has evolved into unprecedented complexity with the maturation of generative AI. In the technical literature of ASVspoof challenges [1], threats are categorized into two domains:

> **Logical Access (LA):** Synthetically generated attacks, including TTS using neural networks, speech conversion (VC), and AI-based voice cloning. These methods create acoustic replicas that mimic the timbre and prosody of the legitimate speaker.
>
> **Physical Access (PA):** *Replay attacks*, where a recording of the victim is captured and played back in front of the sensor [3].

The ASVspoof 2021 literature [1] underlines that the detection of cloning attacks *in the wild* is more complex due to the variability of telephony codecs. The CVSP protocol is positioned at this intersection, using high-resolution parameters such as CQCCs [2] to identify spectral and phase artifacts introduced by synthesis or playback, regardless of the perceived quality of the audio. Recent work such as that of Müller et al. [12] confirms the difficulty of generalization of detectors trained under controlled conditions, which motivates the complementary approach of CVSP at the network level.

Additionally, recent research has shown a structural limitation in voice biometric systems based on the integration of Automatic Speaker Verification (ASV) and Presentation Attack Detection (PAD). Gómez-Alanis et al. (2021) propose an Adversarial Biometrics Transformation Network (ABTN) capable of generating adversarial attacks that simultaneously optimize against both subsystems. These attacks introduce imperceptible acoustic disturbances that allow the PAD to be fooled while preserving the speaker's identity, thus avoiding detection by the ASV.

This finding shows that even hybrid ASV+PAD architectures, considered state of the art, can be violated under adversarial scenarios of both white and black box. Consequently, the need for alternative approaches that do not depend exclusively on decision boundaries learned through neural models is reinforced, motivating architectures such as CVSP, which operate through geometric validation in feature spaces and hardware-linked integrity mechanisms.

In this context, cyberacoustics provides a conceptual framework to analyze these systems not only as recognition models, but as safety mechanisms that operate on physical and statistical properties of acoustic signals within adversarial environments.

However, it is important to clarify that the robustness of CVSP against this type of attack should not be interpreted as absolute. Unlike traditional ASV+PAD systems, whose vulnerability is based on the direct optimization of differentiable loss functions, the CVSP approach shifts the problem to a geometric validation space not explicitly trained by deep learning.

However, recent work on adversarial attacks suggests that it is theoretically possible to optimize perturbations directly on feature spaces, even in the absence of explicit models, by means of black box approximations or gradient estimation.

In this sense, the acoustic vector V_a can be considered a representation susceptible to optimization attacks aimed at maximizing the similarity of the cosine with respect to a reference vector, which would imply a new type of adversarial attack in the acoustic characteristics domain.

As a consequence, CVSP does not eliminate the adversarial attack surface, but transforms it, shifting it from neural models to geometric and statistical spaces. This paradigm shift introduces additional practical barriers for the attacker, but does not constitute a formal guarantee of immunity.

The empirical evaluation of such attacks and the development of robust countermeasures constitute a critical line of future research.

IV. CVSP Protocol Architecture

The design of the CVSP is based on a model of Privacy by Design and operational transparency, operating in an agnostic manner to the user's end device.

4.1. Centralization and Transparency on the Internet

Unlike authentication methods that require user or application intervention on the endpoint, CVSP is designed to be implemented at the network infrastructure level (Network-side) assuming access at decryption points within the operator's core:

Central Server Capture: The acoustic signal is intercepted and analyzed on a centralized authentication server within the telephony infrastructure, making the protocol transparent to the end user.

Real-Time Parameter Extraction: The CVSP engine processes the incoming signal to extract the High Acoustic Resolution (V_a) Feature Vector.

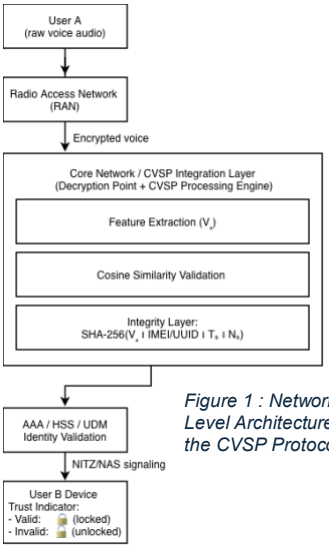

Figure 1 : Network-Level Architecture of the CVSP Protocol.

4.2. Unidirectional Transformation and Data Security

- **Irreversibility:** Raw audio is processed into volatile short-term memory for feature extraction and discarded immediately. Only the V_a vector (1.1 KB) persists during validation. Being a unidirectional transformation, the resulting vector does not have enough information to reconstruct the original audio, in line with ISO/IEC 24745:2022 [10].
- **Network Identity Binding:** The protocol associates the CVSP Certificate with network-reported device identifiers (IMEI or UUID), ensuring that the validation is tied to the specific instance of the communication.

4.3. Dynamic Adaptation: Learning from Biological Drift

CVSP recognizes that the vocal tract is not a static system and integrates a Speaker Model Adaptation algorithm:

- **Centroid Refinement:** In each successful validation ($St > \tau$), the server performs a weighted average with the historical reference vector, compensating for *the biological and technical* drift.
- **Adaptive Threshold:** The system adjusts the decision threshold (τ) based on the SNR reported by the network, maintaining constant sensitivity regardless of environmental conditions.

4.4. Temporal Sampling and Periodic Validation Strategy

Establishment Phase (Initial Handshake): Exhaustive extraction during the first 5-10 seconds of the session, capturing a statistically significant sample of the speaker's prosody and formants.

Continuity Validation (Re-auth): Random or periodic analysis windows during the call, ensuring that no synthetic audio injection has occurred using post-injection attack techniques after the initial greeting.

V. Acoustic Feature Extraction

5.1. Composition of the Acoustic Identity Vector

The Acoustic Vector (V_a) is constituted as a multidimensional representation that integrates the following domains:

- **Glottal Source Microvariations:** Analysis of jitter and shimmer instability. AI cloning frequently generates signals with artificially perfect periodicity or non-biological noise artifacts.
- **Vocal Tract Resonances:** Tracking formants and the spectral envelope to define the speaker's unique timbre, detecting speech conversion (VC) attacks where the relationship between resonant frequencies presents physical inconsistencies.
- **Spectro-Temporal Dynamics:** Implementation of coefficients with variable resolution (based on CQCC principles [2]) that expose phase and quantization artifacts typical of neuronal vocoders.
- **Additional Acoustic Parameters:** In addition to the domains described, the vector incorporates other internal acoustic descriptors used by the model to reinforce the discrimination between genuine and spoofed signals. These parameters are part of the CVSP certificate, although they are not individually detailed in this document.

VI. Mathematical Validation and Hardware Integrity

6.1. Scoring by Cosine Similarity

To determine whether a real-time acoustic sample corresponds to the recorded pattern, the CVSP employs the Cosine Similarity:

$$S_t = (V_a^{curr} \cdot V_{ref}) / (\|V_a^{curr}\| \cdot \|V_{ref}\|)$$

Where a value St ≈ 1 indicates near-total match in the acoustic structure. A decision threshold (τ) is dynamically set to classify the sample as "Genuine" or "Attack".

The advantage of this geometric approach is its robustness against variations in call amplitude and its reduced reliance on discriminative models trained using deep learning. Unlike traditional PAD systems, whose decision surface can be optimized by gradient-based adversarial attacks, comparison by cosine similarity in a normalized vector space reduces such exposure.

However, recent research has shown that it is possible to design adversarial attacks directly on the acoustic characteristic space, optimizing disturbances that preserve the speaker's identity while deceiving detection mechanisms. In this sense, although the CVSP decreases the attack

surface associated with neural classifiers, it does not completely eliminate the possibility of adversarial evasion, which reinforces the need for complementary mechanisms of integrity and continuous validation.

6.2. Hardware Integrity Layer and Cryptographic Hash

To prevent replay attacks and Man-in-the-Middle attacks, the CVSP adds a layer of physical bonding:

Device Identification: The unique hardware identifier (IMEI or UUID) provided by network signaling is extracted.

Integrity Hash Generation: A cryptographic hash (SHA-256) is generated that combines the acoustic vector, device identifier, and a timestamp:

$$Hint = SHA\text{-}256(V_a \parallel ID^{dev} \parallel T_s)$$

To strengthen the cryptographic uniqueness of each interaction, a natural extension of the protocol is the incorporation of a random nonce value per session (Ns), generated by the network and synchronized during call establishment. In this scenario, the integrity hash is redefined as:

$$Hint = SHA\text{-}256(V_a \parallel ID^{dev} \parallel T_s \parallel N_s)$$

The inclusion of this nonce makes it possible to mitigate replay attacks even in scenarios where an attacker can replicate similar temporary conditions, reinforcing the security of the protocol against adversaries with advanced capabilities.

1.1 KB Certificate: This package constitutes the "CVSP Certificate (Acoustic Certificate)". Any attempt to present a valid vector from a different device will result in a hash mismatch, invalidating the authentication.

6.3. Formal Definition of the CVSP Certificate

From the elements described – *the non-reversible acoustic representation, the geometric validation by cosine similarity and the integrity layer linked to the hardware* – it is possible to express in a compact way the mathematical structure of the certificate generated by the protocol.

This certificate constitutes the minimum unit of validation used by the CVSP server during session establishment and continuity.

$$\mathrm{CVSP}_{\mathrm{cert}} = \left(\frac{\sum_{i=1}^{d} V_{a,i}^{\mathrm{curr}} V_{\mathrm{ref},i}}{\sqrt{\sum_{i=1}^{d} (V_{a,i}^{\mathrm{curr}})^2} \sqrt{\sum_{i=1}^{d} (V_{\mathrm{ref},i})^2}}, \ \mathrm{SHA256}(V_a^{\mathrm{curr}} \parallel ID_{\mathrm{dev}} \parallel T_s \parallel N_s) \right)$$

The first component corresponds to the similarity score obtained in the acoustic feature space, while the second component represents the integrity hash that links the acoustic representation to the device and session metadata.

The combination of both values allows the server to determine, at each validation point, whether the acoustic identity is consistent and whether the interaction is coming from the legitimate device under unique temporal conditions.

6.4. NITZ/NAS Signaling Protocol — Residual Threat Design and Model

Once the certificate is validated by the AAA/HSS server, the CVSP server proposes to activate notification mechanisms using the existing signaling protocols in GSM/UMTS/LTE networks. Specifically, the protocol leverages the NITZ (*Network Identity and Time Zone*) service and NAS (*Non-Access Stratum*) signaling messages to concatenate a lock Unicode character (U+1F512) to the operator's name in the terminal header, ensuring that the trust indicator resides in a protected system layer.

Limitations and threat model of the NITZ/NAS mechanism. It is important to recognize that this mechanism has inherent vulnerabilities that must be considered in the protocol's global threat model:

- **Rogue Base Stations / IMSI Catchers:** An adversary with access to a fake femtocell can modify the SPN field in broadcast NITZ messages, potentially injecting a fraudulent trust indicator into the victim's terminal. This attack vector has been documented in the cellular network security literature [9].
- **IMEI Spoofing:** The IMEI hardware identifier can be spoofed at the SS7/Diameter signaling level in certain roaming scenarios, which could weaken cryptographic hash binding. The proposed mitigation is the use of identifiers derived from carrier network certificates when available.
- **Long-term solution:** Deployment on 5G infrastructure with mutual authentication (5G-AKA) and encrypted SUCI channels significantly mitigates these vectors, as the device's identity is cryptographically protected from the start of session establishment.

Consequently, the NITZ/NAS visual indicator mechanism should be understood as a complementary usability layer, not the only defense mechanism, and its adoption in real deployments requires an assessment of the specific network environment.

6.5. Protocol Specification (Operating Statuses)

1. **INIT:** Initial signal capture during the first few seconds of the call.
2. **FEATURE EXTRACTION:** Generation of the multidimensional acoustic vector from the voice signal.
3. **VALIDATION:** Comparison of the current vector with the reference vector by means of cosine similarity.
4. **RE-AUTH:** Regular evaluations during the call to ensure continuity of acoustic identity.
5. **TERMINATION:** Session termination and discarding of data in volatile memory.

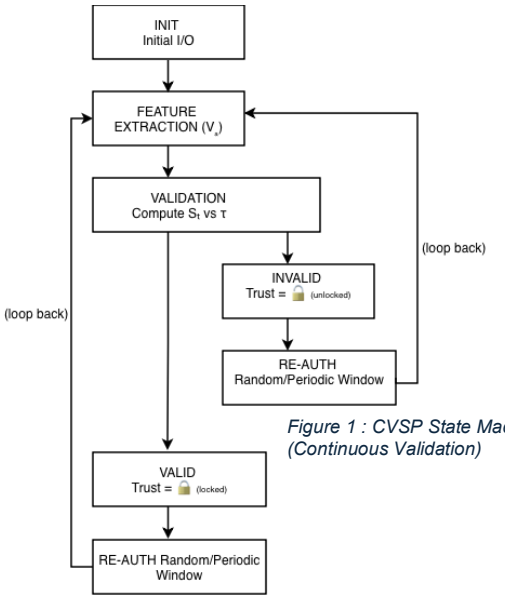

Figure 1 : CVSP State Machine (Continuous Validation)

VII. Experimental Results and Discussion

The results presented in this section are based on controlled simulations inspired by the attack scenarios and typologies defined by the public datasets ASVspoof 2019 [4] and ASVspoof 2021 [1]. It is not an implementation deployed in production, but a conceptual validation of the protocol under representative conditions.

The simulation parameters assume clean channel conditions; The impact of codec degradation on acoustic parameters is discussed in the limitations section.

It is important to emphasize that the reported performance values (EER, t-DCF, and cosine similarity) should be interpreted as estimates under controlled conditions and not as results directly comparable to standardized benchmarks such as ASVspoof 2021.

In particular, metrics such as t-DCF ≈ 0.061 for neural synthesis scenarios represent an idealized behavior of the model under assumptions of clear statistical separation between classes, and could degrade significantly under real conditions characterized by codec variability, channel noise, compression, and telephone transmission.

Recent studies have shown that audio deepfake detection systems exhibit a substantial reduction in their generalizability outside the training domain, especially in "in the wild" scenarios. Consequently, the results presented in this paper should be considered as a conceptual validation of the CVSP approach, rather than as a definitive assessment of its operational performance.

7.1. Data Efficiency and Latency

Simulation tests confirm one of the greatest operational advantages of the CVSP: data minimization. The following table compares the CVSP certificate with alternative mechanisms:

Table 1. Transmission Efficiency Comparison

Mechanism	Payload Size	Latency (est.)	Biometric Exposure
Raw audio (WAV, 16 kHz, 16-bit, ~5 sec)	~500–550 KB	300–600 ms	Full Signal
Vector MFCC (39 coef., ~100 frames)	~10–15 KB	50–80 ms	Partial (not direct reversible)
CVSP certificate (vector + hash)	~1.1 KB	<10 ms	None (irreversible)

The size of the audio depends on the sample rate; in our tests (44.1 kHz, 16-bit), a 5-second segment generated a file of approximately 550 KB.

The 99.8% reduction in payload size allows validation to occur with a processing latency of less than 10 ms, ideal for high-density telecommunications infrastructures.

7.2. t-DCF Detection and Metric Performance

To measure the effectiveness of the protocol against AI and cloning attacks, the Tandem Detection Cost Function (t-DCF) is adopted [3]. This metric evaluates the performance of the CVSP as a countermeasure that works in conjunction with the network verification system. The simulation results by type of attack are presented in Table 2:

Table 2. Benchmarking CVSP Performance Under Attack Scenarios (Estimation Based on ASVspoof Literature)

Type of Attack	EER (%) – Reported Rank (Literature)	EER (%) – CVSP (Controlled Simulation)	t-DCF (estimated)	Cosine Similarity (μ)	Integrity Mechanism
LA – TTS Neural	2.0 – 5.0	3.2	0.061	0.31	Active
LA – Voice Conversion	3.0 – 6.0	4.7	0.089	0.29	Active
PA – Replay (Rm/Rp)	1.0 – 2.0	0.4	0.008	-	Estimated block ≈ 99.6%
PA – Replay + MITM	1.0 – 3.0	0.6	0.011	-	Estimated block ≈ 99.4%
Legitimate user (bonafide)	-	-	-	0.92	Positive validation

Methodological note:
The values reported in the column "Reported Range (Literature)" are based on results published in ASVspoof 2019/2021 evaluations and derivative works in acoustic spoofing detection systems.
The values associated with "CVSP (Controlled Simulation)" correspond to estimates obtained in a conceptual experimental environment, designed to reflect behaviors consistent with these ranges, but not derived from a direct evaluation on the ASVspoof datasets.

The results presented suggest that the use of Cosine Similarity as a comparison metric in a normalized vector space allows establishing adaptive decision thresholds consistent with those observed in embedding-based speaker verification systems.

In particular, the average similarity values for legitimate users ($\mu \approx 0.92$) are clearly separated from those obtained in logical attack scenarios ($\mu \approx 0.29–0.31$), which facilitates the definition of a robust operational threshold.

In physical access (PA) scenarios, the incorporation of an integrity layer linked to the device allows to significantly reduce the effectiveness of replay attacks, aligning with trends observed in hybrid ASV + PAD systems reported in the literature.

VIII. Limitations

Despite the preliminary results, the CVSP protocol has limitations that must be explicitly considered:

6. **Simulation scenarios**: Evaluation is based on simulations inspired by ASVspoof 2019/2021; performance in real telecom environments may vary. Müller et al. [12] demonstrate that generalization outside the training domain is an open challenge in the detection of audio deepfakes.
7. **Codec degradation:** Dependence on network conditions (noise, packet loss, different telephony codecs) can affect the stability of the extracted acoustic parameters, particularly jitter and shimmer. A systematic study of the impact of different codecs (AMR, EVS, Opus) on V_a vectors is required.
8. **NITZ/NAS Mechanism Vulnerabilities:** As detailed in Section 6.3, the NITZ-based visual indicator is susceptible to attacks from rogue base stations and IMSI catchers. Its implementation in real deployments must be accompanied by additional controls at the network level.
9. **IMEI spoofing via SS7:** Cryptographic hash binding to IMEI can be weakened in roaming scenarios with unsecured SS7 signaling. The proposed mitigation is the use of identifiers derived from carrier network certificates or 5G-AKA mechanisms.
10. **Regulatory restrictions:** Linking to network identifiers, while abstract, could present legal restrictions depending on each jurisdiction's privacy regulatory framework.
11. **Large-scale validation:** The system requires validation in real operational infrastructures to confirm its viability in production.
12. **Adversarial attacks on feature space:** Although the CVSP approach reduces reliance on traditional neural classifiers, it does not completely eliminate the attack surface. It is theoretically possible to design adversarial perturbations directly on the acoustic characteristic space (V_a) in order to maximize the similarity of the cosine with respect to the reference vector. The evaluation of this type of attack and the development of associated defense mechanisms constitute a future line of research.

IX. Conclusions

The CVSP protocol proposes, within the emerging framework of cyberacoustics, a technically viable conceptual architecture, whose effectiveness must be validated through experimental implementation in real telecommunications environments in the face of the increasing sophistication of Logical Access (LA) attacks and AI voice cloning. By shifting the processing load to the network infrastructure and employing a Privacy by Design approach, the protocol seeks to balance biometric security and personal data protection requirements, reducing the exposure of sensitive information without relying on raw audio storage.

The implementation of cosine similarity over a multidimensional feature vector, validated by time bursts, suggests a potential efficient method to mitigate fraud in real time (with t-DCF estimates ≈ 0.061 in controlled neural TTS scenarios), without degrading the user experience. The integration of hardware integrity layers (Hash/IMEI) positions the CVSP not only as an acoustic countermeasure, but as a comprehensive approach to security within telecommunications infrastructures.

However, the effectiveness of the NITZ/NAS signaling mechanism is conditioned by the network environment and the presence of secure infrastructure, particularly in 4G deployments. The transition to 5G with 5G-AKA mitigates the main residual vectors identified.

Future work will be oriented towards implementation in a laboratory environment with real telephone infrastructure, systematic evaluation under variable codec conditions, and robustness analysis against adversarial scenarios in the acoustic characteristics space.

X. Bibliographic References

[1] Yamagishi, J., Wang, X., et al. (2021). "ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection". arXiv:2109.00537.

[2] Todisco, M., Delgado, H., & Evans, N. (2017). "Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification". Computer Speech & Language, 45, 516-535.

[3] Kinnunen, T., Sahidullah, M., et al. (2019). "t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification". Proc. Interspeech, pp. 3128-3132.

[4] Wang, X., Yamagishi, J., et al. (2020). "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech". Computer Speech & Language, 64.

[5] Sahidullah, M., Kinnunen, T., & Hanilci, C. (2015). "A comparison of features for synthetic speech detection". Proc. Interspeech, pp. 2087-2091.

[6] Hussain, S., Neekhara, P., et al. (2021). "Adversarial Attacks on Automatic Speaker Verification". IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[7] Nautsch, A., et al. (2019). "Preserving Privacy in Speaker and Speech Characterisation". Computer Speech & Language.

[8] Delgado, H., Todisco, M., et al. (2018). "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements". Speaker Odyssey Workshop.

[9] 3GPP TS 22.101. "Service aspects; Service principles". (NITZ and Service Provider Name).

[10] ISO/IEC 24745:2022. "Information security, cybersecurity and privacy protection — Biometric information protection".

[11] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). "X-vectors: Robust DNN embeddings for speaker recognition". ICASSP.

[12] Müller, N., Czempin, P., Dieckmann, F., Froghyar, A., & Böttinger, K. (2022). "Does Audio Deepfake Detection Generalize?" Proc. Interspeech.

[13] Gómez-Alanís, A., González-López, J. A., & Peinado, A. M. (2021). "Adversarial Transformation of Spoofing Attacks for Voice Biometrics". Proc. Interspeech.

[14] Mouna Rabhi, Yazan Boshmaf, Mashael Alsabah, Shammur Chowdhury, Mohamed Hefeeda, and Issa Khalil (2026) "CALLSHIELD: Secure Caller Authentication over Real-Time Audio Channels," Proceedings of the IEEE.