RIRplay: Generation of a Replay Stereo Corpus for Voice Biometrics Anti-Spoofing

Jose C. Sanchez-Valera, Antonio M. Peinado, *Senior Member, IEEE*, Juan M. Martin-Doñas, Alejandro Gomez-Alanis, Angel M. Gomez, and Massimiliano Todisco, *Member, IEEE*

***Abstract*—While recent efforts in countering spoofing attacks on voice biometric systems have primarily focused on detecting synthetic speech, Physical Access (PA) attacks, such as audio replay, still pose a serious and unresolved challenge. This research gap has been mainly due to the lack of new, realistic speech corpora for training and testing effective and generalizable countermeasure systems. Given the difficulty in collecting actual audio samples from this kind of attack, simulation has been proposed as an alternative to provide audio replay training data. The objective of this work is the generation of a novel simulated database, called RIRplay, that is both realistic, in the sense of reproducing the actual spoofing process, and representative of a wide variety of possible acoustic contexts. Our results show that training with the RIRplay corpus reduces the Equal Error Rate (EER) by nearly 10 percentage points on the challenging ASVspoof 2021 evaluation set, from 36.89% to 28.04%, compared to models trained on the ASVspoof 2019 corpus, demonstrating significant improvements in out-of-domain generalization.**

***Index Terms*—Anti-spoofing, voice biometrics, replayed audio, physical access, simulated database.**

I. Introduction

THE increasing digitalization of services requires secure authentication methods to prevent unauthorized access [1]. Biometric systems fulfill this purpose by matching discriminatory attributes of acquired signals (test signals) with stored signals (enrollment signals) that uniquely identify users [2]. For speech signals, this procedure is known as Automatic Speaker Verification (ASV), which compares a user's voice against an enrolled profile, either text-dependent (TD) or text-independent (TI) to confirm identity [3], [4], [5]. In general, these ASV systems (especially those operating in TI mode [6]) are vulnerable to spoofing attacks. These include (pre-recorded) replayed speech and synthetic voices from text-to-speech (TTS) or voice conversion (VC) systems, which may also be replayed in order to hide their synthetic origin [7].

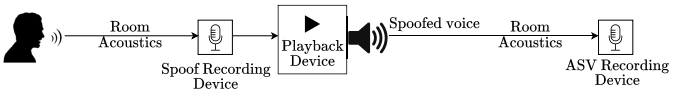


Fig. 1. Operation scheme of the replay attack.

Among the different attacks that may be crafted to deceive an ASV system, those based on surreptitious recordings from real speakers, which are later replayed through a loudspeaker to impersonate them, as illustrated in Figure 1, must be particularly taken into account. The reason is that they do not require any technical knowledge on the attacker's side, making them the most accessible and easy-to-perform [8], [9], [10].

Despite the simplicity of performing a replay attack, acquiring a real speech corpus for training and testing replay anti-spoofing systems requires a cumbersome experimental setup. This involves recording multiple real speakers in different acoustic environments, using ASV and spoofing microphones placed at different locations, and employing various loudspeakers to reproduce the speech signal. This complex procedure makes the generation of real replay speech signals neither appealing nor practical. Thus, generating a corpus both realistic and capable of generalizing to different situations, remains an unsolved task [11].

In this work, we present a novel simulation method to create a replay speech database for anti-spoofing in voice biometrics, following a data-centric approach that prioritizes creating diverse and realistic training examples rather than increasing the complexity of the classification model. By modeling key acoustic elements–room impulse responses (RIRs), noise, and loudspeaker characteristics–our method closely mimics real spoofing conditions, enhancing generalization to unseen data domains. Compared to ASVspoof 2019 Physical Access (PA) [12], [13], our corpus, hereinafter referred to as RIRplay, better reflects real-world scenarios and includes a stereo pipeline linking each spoofed signal to its bona fide source. Moreover, both the proposed generation method and the resulting corpus can be useful for emerging fields of research on replayed deepfake attacks [7].

The rest of this paper is organized as follows. Section II analyzes the limitations of existing PA replay datasets. Section III outlines the proposed data generation framework, with

Received 6 May 2025; revised 29 December 2025 and 20 February 2026; accepted 1 April 2026. Date of publication 16 April 2026; date of current version 23 April 2026. This paper is part of the project PID2022-138711OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU. The associate editor coordinating the review of this article and approving it for publication was Prof. Paolo Bestagini. *(Corresponding author: Jose C. Sanchez-Valera.)*

Jose C. Sanchez-Valera, Antonio M. Peinado, and Angel M. Gomez are with the Department of Signal Theory, Telematics and Communications, University of Granada, 18071 Granada, Spain (e-mail: svjosecarlos@ugr.es; amp@ugr.es; amgg@ugr.es).

Juan M. Martin-Doñas is with the Department of Industrial Engineering, Universidad de La Laguna, 38200 La Laguna, Spain (e-mail: jmartido@ull.edu.es).

Alejandro Gomez-Alanis is with the Amazon AGI Labs, 52062 Aachen, Germany (e-mail: agomezalanis@ugr.es).

Massimiliano Todisco is with the Audio Security and Privacy Group, EURECOM, 06904 Biot, France (e-mail: massimiliano.todisco@eurecom.fr).

Digital Object Identifier 10.1109/TIFS.2026.3684815

TABLE I
CURRENT LIMITATIONS IN THE ASVSPOOF 2019 PHYSICAL ACCESS (PA) DATABASE AND PROPOSED ENHANCEMENTS

Aspect	Limitations	Proposed modifications
Room acoustics realism	RT60 is kept constant, but materials have varying absorption coefficients across frequency [14].	We will use real room parameters with experimentally measured RT60 values for each octave band.
Physical situations	Some acoustic combinations are not entirely realistic and reflect impossible situations (e.g., a room size of $2\,\mathrm{m}^2$ with a reverberation time (RT60) of $1000\,\mathrm{ms}$).	Acoustic measures of real rooms, in which the reverberation time, distances and sizes are consistent with each other, will be adopted.
Real-world noise	The audios exclude environmental noise, failing to replicate real conditions where noise occurs during spoofed forgeries.	All audios must involve real, experimentally recorded additive noises, such as babble and other realistic sounds like general urban background noise.
Subject size and location	According to [12], the microphones' and speakers' heights are fixed at $1.1\,\mathrm{m}$, keeping the elevation angle at 0° and sound propagation constant.	The positions and sizes of the speaker and attacker should be as random as possible, ensuring varied elevation angles and propagation paths.
Class silence duration	Bhusan Chettri *et al.* [15] found that silence duration varies between bona fide and spoofed classes, which may serve as a misleading shortcut for classification.	The bona fide-spoofed audio generation process must be refined by compensating for convolution-induced delays and trimming leading and trailing silences.

Section IV detailing its implementation. Section V describes additional refinements, such as silence suppression and greater data diversity. Section VI details the experimental setup and presents the results, demonstrating the dataset's effectiveness. Finally, Section VII concludes with a summary of key findings.

II. RELATED WORK

Replay speech databases have been crucial for advancing voice anti-spoofing research. Over the past decade, several datasets have been introduced, including ReMASC [16], RedDots Replayed [17], LRPD [18], VSDC [19], and those from the ASVspoof challenge series [20], such as ASVspoof 2017 [21] and ASVspoof 2019 [12] datasets. This section reviews these key resources, noting their limitations on PA detection.

A. Limitations of Existing Realistic Corpora

Generating realistic datasets that replicate physical replay attacks is a natural solution but faces several issues, including limited capture and playback devices, a small number of speakers, and restricted acoustic diversity, which may lead to poor performance due to insufficient acoustic variability [22], [23]. For example, **ReMASC** [16], recorded in real indoor and outdoor environments, is limited to four locations and five loudspeakers, while **RedDots Replayed** [17] has recordings from only four locations. Although **VSDC** [19] increases the number of environments, it still offers limited attack diversity with up to fourteen playback devices and only fifteen speakers.

Other datasets aim to cover a broader range of environments and attack configurations. For example, **ASVspoof 2017** [21] includes 26 environments, 25 recording devices, and 26 playback loudspeakers, while the **LRPD** database [18] features a larger speaker population and diverse device and environmental settings. However, this increased diversity comes at the cost of oversimplified replay scenarios. These datasets rely on publicly available speech corpora as bona fide trials or as direct inputs to the playback-recording pipeline, assuming that attackers have direct access to the target speakers' audio files and omitting the attacker's capture stage, while also ignoring the effects of the presentation environment (i.e., room and ASV system) on bona fide accesses. In contrast to the complete spoofing chain in Figure 1, they further omit cascading room-acoustic and playback-recording effects. These simplifications stem from the high cost and complexity of collecting realistic replay data across diverse speakers, devices, and locations.

Finally, it is worth mentioning that these datasets often present clear biases and shortcut design artifacts between bona fide and spoofed recordings that restrict generalization, causing models to rely on spurious cues rather than meaningful discriminative information. A notable example is ASVspoof 2017, which contains artifacts such as initial non-speech noise, burst clicks, and corrupted files [24]. Although some of these issues can be partially corrected, they are difficult and costly to fully resolve once a database has been released.

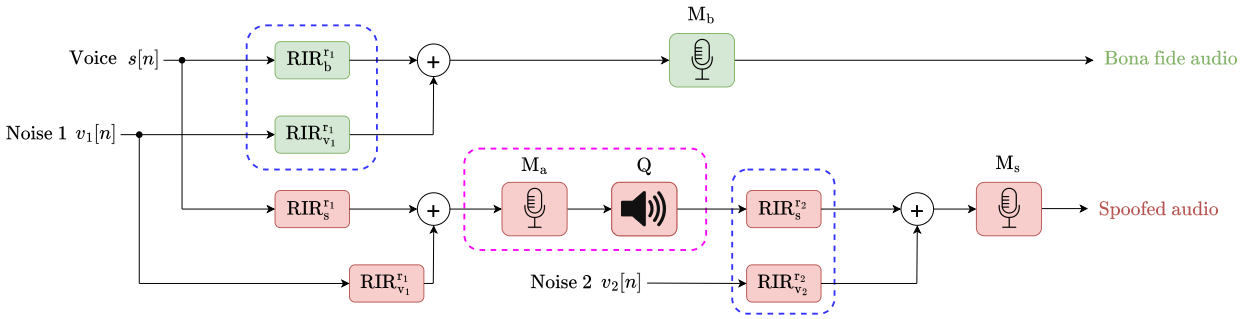
The above examples clearly reflect the challenges of designing realistic spoofing corpora and motivate the need for novel design approaches to improve detection generalization.

B. ASVspoof 2019: A Simulated Solution

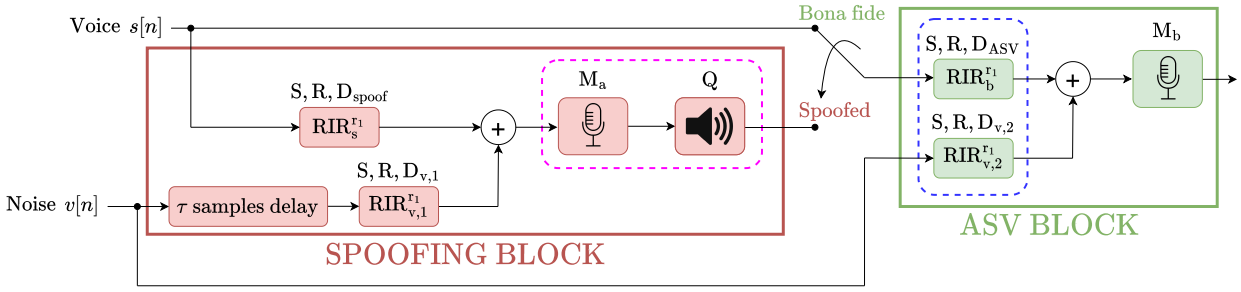
Unlike earlier corpora attempting to emulate real replay attacks, ASVspoof 2019 [12] uses a simulated approach, enabling greater variability in environments, speakers, and devices. This allows a controlled and reproducible experimental setup suitable for systematic evaluation. Despite its effectiveness [13], models trained on ASVspoof 2019 exhibit a notable performance drop on real evaluation benchmarks like ASVspoof 2021 [22], which relies on physically replayed recordings. This degradation evinces a mismatch between simulated and real data which eventually reduces models' ability to generalize to unseen real contexts. Table I (Column 2) enumerates some ASVspoof 2019 design aspects which either lack realism or may reduce detection performance.

III. REPLAY ATTACK SIMULATION METHODOLOGY

To overcome ASVspoof 2019's limitations, the simulation should more closely mimic the actual environments where genuine speech and replay attacks occur. We address this with RIRplay, a new simulated corpus developed under more realistic assumptions while also providing acoustic diversity. Specific modifications are summarized in Table I, Column 3.



(a) Fully realistic simulation pipeline.



(b) Proposed generation method.

Fig. 2. Block diagrams of (a) the fully realistic simulation pipeline and (b) the proposed method. The blocks $\text{RIR}_{\text{b}}^{\text{r}_1}$, $\text{RIR}_{\text{s}}^{\text{r}_1}$, and $\text{RIR}_{\text{s}}^{\text{r}_2}$ denote the RIRs applied to the bona fide signal before reaching the ASV microphone (M_b), as well as to the attacker's original recording and replay stages, respectively. $\text{RIR}_{\text{v}_1}^{\text{r}_1}$ and $\text{RIR}_{\text{v}_2}^{\text{r}_2}$ represent the RIRs affecting environmental noise during speech capture and replay in the generic scenario, whereas $\text{RIR}_{\text{v,1}}^{\text{r}_1}$ and $\text{RIR}_{\text{v,2}}^{\text{r}_1}$ are used for spoofed recording noise and ASV presentation in our proposal. Parameters Q, M_a, and M_s denote the loudspeaker and the attacker's capture and ASV presentation microphones, respectively. Finally, S, R, and $\text{D}(\cdot)$ indicate room size, reverberation time, and the distances from sources to microphones.

TABLE II
ASVspoof 2019 Distribution for Training, Development and Evaluation Sets

	Training	Development		Evaluation	
		Target	Non-target	Target	Non-target
Bona fide	5,400	2,700	2,700	12,960	5,130
Spoofed	48,600	24,300		116,640	
Total	54,000	29,700		134,730	

TABLE III
Number of VCTK Utterances Used to Build the Database

Training set	Development set		Evaluation set	
	Target	Non-target	Target	Non-target
200 utts	100 utts	100 utts	480 utts	190 utts

A. Constraints

For several years, PA anti-spoofing research focused on improving countermeasure (CM) systems, neglecting the importance of having realistic and good quality data for training. The proposed data generation methodology addresses this gap while retaining some structural features of ASVspoof 2019:

1) Enabling a fair comparison between this new database and the original ASVspoof 2019 dataset. As shown in Table II, we will retain the same data amounts as in ASVspoof 2019 to ensure consistency in benchmarking. Additionally, to enhance generalization, we also consider an extended version of our RIRplay database, increasing the amount of utterances to improve model robustness. This setup resembles the benefits of data augmentation (DA) techniques [23], [25], [26], [27], offering a broader data diversity.
2) Preserving the phonetic and vocal features of the different speakers in the ASVspoof 2019 database. To accomplish this, we employ the same seed utterances (from the same speakers) from the VCTK Audio Corpus [28], as used in ASVspoof 2019. Table III reports the number of original utterances employed in each subset.

B. Methodology

To generate a realistic simulated dataset, we first identify the acoustic elements involved in signal acquisition. Figure 2a outlines the different steps involved in a generic PA scenario, in which the speaker's speech is captured, during a genuine access trial, either by the ASV microphone (M_b) or by the attacker's microphone (M_a), and later replayed through a loudspeaker (Q) to spoof the ASV system via M_s. Both ASV recordings may occur in completely different rooms ($\text{r}_1 \neq \text{r}_2$), also involving different noises ($v_1[n] \neq v_2[n]$), and different ASV microphones ($\text{M}_\text{b} \neq \text{M}_\text{s}$), or within the same room ($\text{r}_1 = \text{r}_2$), which represents the most challenging scenario [12] (see Section 2.2.1). In this latter configuration, the replay attack closely replicates the genuine access trial of the speaker:

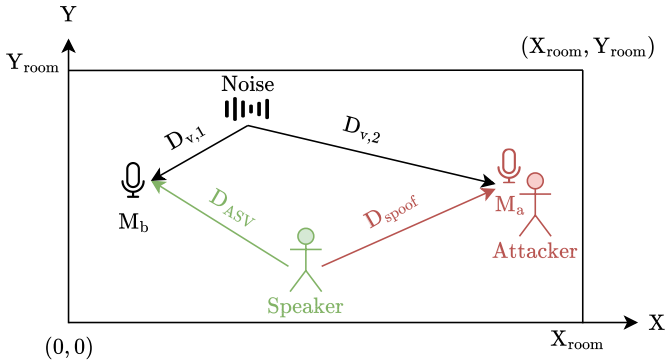


Fig. 3. Most challenging scenario for anti-spoofing detection. The attacker is in the same room as the genuine speaker.

TABLE IV

SIZE AND AVERAGED RT60 OF THE CHARACTERIZED ROOMS IN [29]. RT60 VALUES ARE AVERAGED ACROSS FREQUENCY

Room name	Size (m)			Av. RT60 (s)	
	X_{room}	Y_{room}	Z_{room}	1	2
Office 1	4.83	3.32	2.95	0.38	0.38
Office 2	5.10	3.22	2.94	0.43	0.43
Building Lobby	5.13	4.47	3.18	0.66	0.77
Meeting Room 1	5.11	6.61	2.95	0.47	0.49
Meeting Room 2	9.07	10.32	2.63	0.43	0.40
Lecture Room 1	9.73	6.93	3.00	0.66	0.69
Lecture Room 2	9.29	13.56	2.94	1.20	1.20

similar noises (a single source $v[n]$ modified by either $RIR_{v,1}^{r_1}$ or $RIR_{v,2}^{r_1}$), same positions within the room (i.e., a single RIR for ASV presentation, $RIR_{b}^{r_1}$), and identical ASV microphones (M_b). To improve the trained detectors, we focus on this worst-case scenario, where the attacker and speaker share the same room, but experience distinct RIRs for the subreptitious recording and the ASV presentation, as shown in Figure 3.

The previous scenario motivates our data generation scheme, illustrated in Figure 2b, which is structured into two main blocks. The first block, designated as **spoofing block**, focuses on spoof-specific transformations, simulating key alterations that occur during a spoofing attack, such as room RIRs, background noise during recording, microphone response, and loudspeaker nonlinearities. The second block or **ASV block**, on the other hand, models the signal acquisition for ASV, which includes new RIRs and additive noise. By isolating spoofing-specific features from ASV-related variability, the dataset emphasizes the essential characteristics of replay attacks, while including acoustic diversity that helps to fight overfitting. As can be observed in Figures 2b and 3, there are different critical parameters affecting RIR generation: room size (S), reverberation time (R), and distances (D_{ASV}, D_{spoof}, D_v). Every (S, R) pair defines an acoustic environment which will be notated as SR. Both ASV and spoofing blocks share the same SR parameters but vary distances. Additionally, the noises affecting ASV and spoofed recordings are treated separately (a delay τ is introduced between them) to reflect the fact that, although arising from the same source, they are acoustic events that occur at different times.

Additionally, we reflect the stereo nature of our approach, in which spoofed and bona fide samples are inherently paired, by explicitly storing this correspondence in the provided metadata of our dataset. This added feature will facilitate future research by enabling the exploitation of the shared information between bona fide and its corresponding spoofed versions.

Finally, it is important to acknowledge the limitations of our approach: while our proposed generation method effectively simulates replay attacks, it only partially emulates the complexity of real environments. Key elements are considered (as listed in Table I), but other potentially important factors, such as non-rectangular room layouts, furniture, etc., are not explicitly modeled. To sum up, our proposal, although still limited in some aspects, represents a step forward towards the generation of more diverse and realistic replay corpora.

IV. DATASET GENERATION PROCEDURE

Audio generation is simulated following the scheme in Figure 2b. For bona fide voices, every original VCTK audio file [28] undergoes the operations of the ASV block, while spoofed audios must be processed first by the spoofing block.

The following subsections detail the RIRplay data generation process. First, Section IV-A describes the simulation of acoustic environments, which is essential for recreating realistic conditions and involves selecting suitable distances among the actors in the replay scenario. Next, Section IV-B addresses noise addition, while Sections IV-C and IV-D focus on modeling replay devices and on the delay compensation and normalization techniques used to ensure consistency between bona fide and spoofed samples. Finally, Sections IV-E and IV-F present the database partitioning strategy and the metadata describing the dataset structure.

A. Acoustic Environments

The generation of realistic RIRs is one of the main issues we have considered to develop our simulated corpus. The different RIRs have been obtained through Roomsimove [30] using a sampling frequency (F_s) of 16 kHz. This software enables RIR simulation using the image method [31] and requires only a few physical parameters, such as source and microphone positions, and SR environments (as defined in Section III-B). Roomsimove simulates the propagation of sound from the sources, speech and noise in our case, to each receiver in the room (spoofing and ASV microphones).

In this work, we use real room parameters obtained from The ACE Challenge [29], which provides experimental measurements of RT60 values across frequency octaves (125 Hz, 250 Hz,..., 8000 Hz) in seven distinct rooms (see Table IV). For each room, two source-microphone configurations were used during the recording process. While the Challenge employed microphones of different types, we specifically use RT60 measurements obtained with the 8-channel microphone array. This results in a total of 112 acoustic environments (7 rooms $\times$2 configurations $\times$8 channels).

The RIR simulation procedure requires not only the RT60 values per octave (depicted in Figure 4) but also the positions of the subjects. For each acoustic environment SR_i, defined by its room size (S) and RT60 vector (R), we generate a set of U random distances between the speaker and the ASV

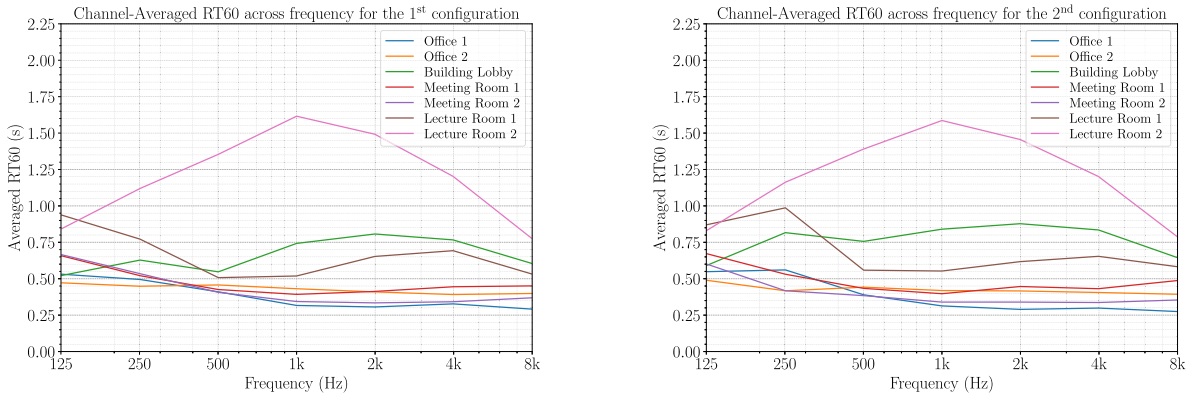


Fig. 4. Channel-Averaged RT60 across frequency for each room characterized in [29]. Subfigures (a) and (b) correspond to the two possible configurations. The reverberation behavior in lecture Room 2 stands out, exhibiting higher variability due to its open structure and reflective surfaces.

TABLE V
DISTANCE-BASED INTERVALS FOR THE ASV SYSTEM AND THE ATTACKER

	Low	Medium	High
$\mathbf{D}_{\mathbf{ASV}}^{i,j}$ **(cm)**	[10, 50]	[50, 100]	[100, 150]
$\mathbf{D}_{\mathbf{spoof}}^{i,j,k}$ **(cm)**	[10, 50]	[50, 100]	[100, 0.5max(X_{room}, Y_{room})]

microphone, denoted as D_{ASV}, i.e., $\{D_{ASV}^{i,j} \mid j = 1, 2, \ldots, U\}$. Additionally, for each of these distances, we define K different values of D_{spoof} (speaker-to-spoofing-mic distance), forming the set $\{D_{spoof}^{i,j,k} \mid k = 1, 2, \ldots, K\}$. This results in a total of $112 \times U \times K$ possible combinations. The classification of the distances into low, medium, or high is detailed in Table V, which specifies the intervals for each category. Notably, the upper limit of $D_{spoof}^{i,j,k}$ is set to 50% of the maximum between the room's width and length. To ensure variability, $D_{ASV}^{i,j}$ and $D_{spoof}^{i,j,k}$ values are randomly selected from the intervals specified in Table V using a uniform distribution.

Regarding the noise sources, for each combination defined by SR_i and $D_{ASV}^{i,j}$, the position of the noise source in the room is randomly chosen. As shown in Figure 3, the resulting distances to the ASV system and the attacker differ (that is, $D_{v,1}^{i,j} \neq D_{v,2}^{i,j}$).

B. Noise: Types and Signal-to-Noise Ratios (SNR)

All additive noise recordings were sourced from [32], covering real-world scenarios including indoor noises (e.g., equipment, conversational babble) and outdoor ambient noises (e.g., traffic and general urban background sound). For each VCTK speech sample $s[n]$, a noise excerpt $v[n]$ is randomly selected. In order to keep acoustic consistency between a bona fide utterance and its corresponding spoofed versions (sharing the same room), the noises applied in the spoofing and ASV blocks (see Figure 2b) are extracted from the same noise recording but they are delayed to reflect the temporal gap between pre-recording and replay. The recording and replaying SNRs (SNR_{rec} and SNR_{rep}) are selected randomly from an uniform distribution in the range [25, 35] dB. In this way, the audios generated have different SNRs, which increases the variability of the data and prevents model overfitting.

In both ASV and spoofing blocks, the addition of the RIR-filtered noise excerpt to the RIR-filtered speech signal, $s_h[n]$, is performed as follows:

$$\tilde{s}_h[n] = s_h[n] + \sqrt{\frac{P_{s_h}}{P_{v_h} \cdot \text{SNR}}} \cdot v_h[n], \tag{1}$$

where $\tilde{s}_h[n]$ represents the final noisy speech signal, P_{s_h} and P_{v_h} denote the power of the speech and noise signals, respectively, and SNR corresponds to either SNR_{rec} or SNR_{rep}, depending on the processing block.

C. Recording and Replaying Devices

To model the non-linear effects introduced by loudspeakers, we simulate their behavior using the Generalized Polynomial Hammerstein Model (GPHM) [33], [34], following the same approach adopted in ASVspoof 2019 [12]. This model generates an output signal $y[n]$ from a given input $\tilde{s}_h[n]$ as

$$y[n] = \sum_{l=1}^{L} h_Q^l[n] * (\tilde{s}_h[n])^l, \quad n \in \mathbb{Z}, \tag{2}$$

where $h_Q^l[n]$ denotes the l-th harmonic impulse response characterizing a given loudspeaker Q, and L is the highest model order. We employ the same set of impulse responses used in ASVspoof 2019 (see [12, Table 5]), which characterize 40 different loudspeakers classified into three quality levels: perfect, high, and low.

In contrast to the loudspeaker modeling, and following the same approach as in [12], our generation method does not model the frequency response of the microphones, as their responses, even in miniaturized devices, are typically linear and relatively flat over the frequency range of interest.

D. Delay Compensation and Audio Normalization

Previous studies have shown that differences in silence content can introduce exploitable artifacts in deep-learning models [15], as observed in ASVspoof 2019 LA [35] and ASVspoof 2017 [24], [36]. These artifacts may cause CM models to overfit and limit generalization to datasets such as ASVspoof 2021. To address this issue, convolution-induced delays are compensated in both ASV and spoofing blocks, to

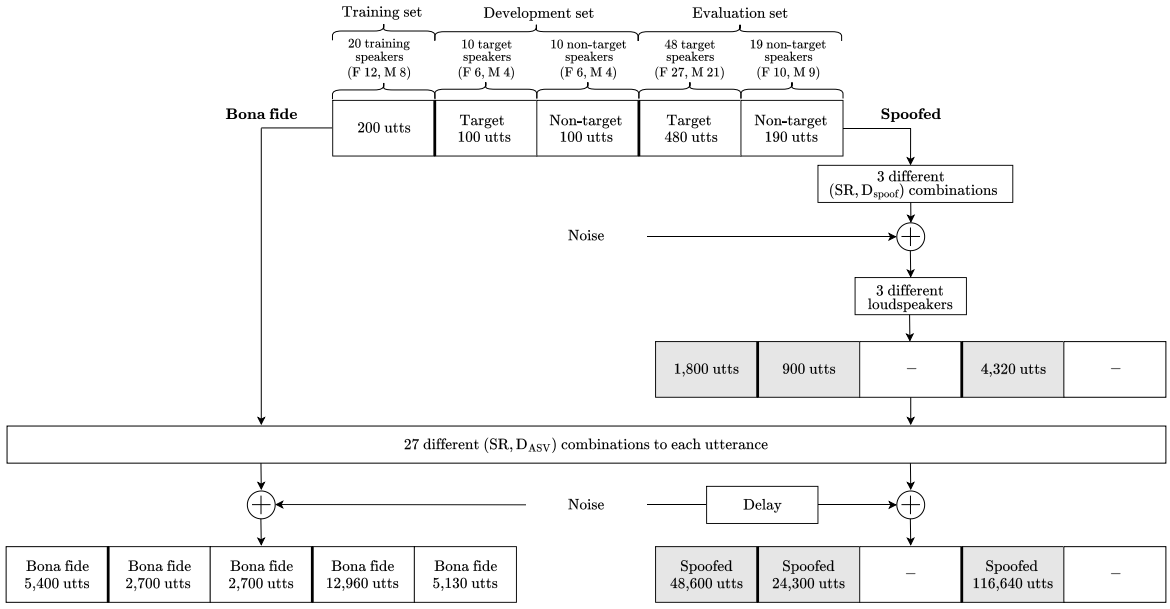


Fig. 5. RIRplay partitions and simulation procedure. Each data block shows the operations and number of utterances in the corresponding subset.

ensure proper alignment between bona fide and spoofed audios in our dataset.

Specifically, since the convolution of the signal $s[n]$ with a RIR $h[n]$, given by $s_{\mathrm{h}}[n] = s[n] * h[n]$, introduces a delay of n_{h} samples, we compensate for by shifting the signal as $s_{\mathrm{h}}^{\mathrm{DC}}[n] = s_{\mathrm{h}}[n + n_{\mathrm{h}}]$, where n_{h} is the position of the maximum value in $h[n]$, typically corresponding to the direct path.

Additionally, arising from the power series associated with the linear filters involved in the GPHM, we observed that the loudspeaker harmonics exhibit a fixed delay of $n_{\mathrm{Q}} = 2048$ samples. Thus, we apply again a delay compensation to the output of Equation (2) (i.e., $y^{\mathrm{DC}}[n] = y[n + n_{\mathrm{Q}}]$).

To conclude, after the entire simulation process, an amplitude normalization is carried out to guarantee that the signal remains within the interval of $[-1,1]$ thus preventing clipping effects and gain-related biases.

E. Database Simulation and Partitioning

In order to generate the custom subsets, the acoustic variability is crucial to avoid overfitting in training and misleading results in evaluation. Consequently, we will distribute the 112 SR acoustic environments from The ACE Challenge as follows:

- Training: 50 random environments are selected.
- Development: 12 additional SR instances are randomly chosen, along with 8 previously seen SR environments.
- Evaluation: the remaining 50 SR samples are used for this dataset.

Figure 5 presents the complete data generation procedure. As shown, one acoustic environment SR_i is randomly selected for each original VCTK audio file [28]. In order to ensure that the number of audios in the database remains equivalent to ASVspoof 2019 [12], every acoustic environment SR_i is combined with $U = 27$ different distances to the ASV microphone (randomly selected from the three categories in Table V, IX per category), thus obtaining a set $\{(\mathrm{SR}_i, \mathrm{D}_{\mathrm{ASV}}^{i,j}) \mid i = 1,2,\ldots,N;\ j = 1,2,\ldots,U\}$ of possible acoustic combinations. Note that the value of N depends on the RIRplay subset and is fixed at 50, 20, and 50 for the training, development, and evaluation subsets, respectively, as specified above. Moreover, the number of spoofed utterances is also preserved by considering $K = 3$ distinct distances (one per category in Table V) to the spoofing microphone, which yields a set $\{(\mathrm{SR}_i, \mathrm{D}_{\mathrm{ASV}}^{i,j}, \mathrm{D}_{\mathrm{spoof}}^{i,j,k}) \mid i = 1,2,\ldots,N;\ j = 1,2,\ldots,U;\ k = 1,2,\ldots,K\}$ of acoustic combinations.

The different loudspeakers are distributed among the training, development, and evaluation subsets as in [12, Table 5]. For each spoofed utterance at distance $\mathrm{D}_{\mathrm{spoof}}^{i,j,k}$, three loudspeakers, one per quality level, are randomly selected to ensure playback diversity, with seen devices for training and development, and unseen ones for the remaining spoofed data in evaluation.

Finally, it must be noticed that the generation scheme of ASVspoof 2019, although providing the same final numbers, is different: the same 27 acoustic environment $(\mathrm{S}, \mathrm{R}, \mathrm{D}_{\mathrm{ASV}})$ categories are applied to every utterance and then combined with 9 different *attacks* $(\mathrm{Q}, \mathrm{D}_{\mathrm{spoof}})$ for spoofed generation. Also, note that all rooms, configurations, and channels from The ACE Challenge may be shared across the three RIRplay subsets. Noise files are also reused, but different excerpts are selected for each utterance as specified in Section IV-B. These design choices are intended to maximize variability within each subset, given that our primary objective is out-of-domain evaluation, i.e., testing on a different database.

F. Metadata Files and Protocols

The audio simulation process is designed to generate paired files, creating different spoofed versions for each bona fide audio. These pairs are systematically documented in a metadata file for each subset, as shown in Table VI.

The file provides protocol information, such as the audio file name (AudioName); the utterance class (Type; bona

TABLE VI
EXAMPLE OF METADATA FILE CREATED DURING THE GENERATION OF BONA FIDE AND SPOOFED AUDIOS IN THE TRAINING SUBSET

AudioName	Type	BonafideName	RoomName	Config	Channel	TypeDasv	TypeDspoof	Dasv	Dspoof	NoiseName	SNRrep	SNRrec	Quality	NLName	isSeen
p229_228_mic1_1000000	bona fide	–	Office1	1	1	0	0	0.41	–	CMC_CT_T	16.92	–	–	–	–
p229_228_mic1_1000001	bona fide	–	Office1	1	1	1	0	0.34	–	CMC_FT_T	19.38	–	–	–	–
p229_228_mic1_1000002	bona fide	–	Office1	1	1	2	0	0.36	–	BAB_CT_T	22.86	–	–	–	–
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
p229_228_mic1_1000027	spoof	p229_228_mic1_1000000	Office1	1	1	0	0	0.41	0.25	CMC_CT_T	16.92	21.22	perfect	S1	S
p229_228_mic1_1000028	spoof	p229_228_mic1_1000000	Office1	1	1	0	0	0.41	0.25	CMC_CT_T	16.92	21.22	high	SS	S
p229_228_mic1_1000029	spoof	p229_228_mic1_1000000	Office1	1	1	0	0	0.41	0.25	CMC_CT_T	16.92	21.22	low	S7	S
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

fide or spoof); the corresponding bona fide audio for each spoof (BonafideName); the name of the acoustic environment (RoomName); the configuration number of The ACE Challenge (Config); the microphone channel (Channel); the distances (TypeDasv, TypeDspoof) along with their metric values (Dasv, Dspoof; in meters); the noise file name (NoiseName); the SNR values (SNRrep, SNRrec; in dB); and, finally, the loudspeaker quality (Quality), its name (NLName), and whether it was seen in training or not (isSeen; S or U).

V. DATASET REFINEMENTS

In this section, we consider a series of improvements intended to increase the generalization capability of CM models trained with the proposed RIRplay dataset. Each improvement targets a specific limitation identified in our data generation pipeline:

1) VCTK utterances may still contain short silences at the start and end of voice recordings. This undesirable phenomenon should be minimized as much as possible, as learning artifacts from silences may lead to a poor generalization capability in cross-domain evaluation. To mitigate this issue, we will apply a Voice Activity Detector (VAD) to all generated audio samples.
2) As discussed in Section II-A, and also as shown in Section VI-C, increased data variability is known to improve generalization and robustness. In order to boost this effect, we will extend the database size by doubling the number of VCTK seed utterances and combining them with new acoustic contexts, thus obtaining a doubled-size corpus.

The implementation details of these improvements are explained in the following subsections.

A. Suppression of Initial and Final Silences

One of the main strengths of our data generation scheme is its ability to easily visualize which artifacts may affect detection when using deep-learning models. As shown in Figure 2b, the expected differences between classes primarily stem from loudspeaker effects and inherent acoustic discrepancies. To ensure the CM system can learn these subtle cues, it is essential to avoid unintended shortcut elements such as silence periods.

Previous work [15] demonstrated that models trained on ASVspoof 2019 tend to overfit due to class-dependent silence patterns, and showed that removing leading and trailing silences improves generalization.

While the delay compensation applied in our approach partially mitigates this issue, it does not fully remove residual silence that may bias the CM system. To overcome this limitation, we apply a Voice Activity Detector (VAD) to eliminate non-speech segments at the beginning and end of recordings. During training and testing we employ sileroVAD [37], and additionally evaluate using rVADfast [38] to ensure that improvements do not depend on the VAD used.

B. Extended Corpus

To increase data diversity by a two-factor, we double the number of seed utterances per speaker from the VCTK corpus. However, while doubling the utterances per speaker increases phonetic diversity, it does not necessarily improve the acoustic variability. To address this, the new generation process involves assigning acoustic environments to unused noise fragments. Thus, replicating the procedure described in the previous section yields a more diverse corpus of twice the size.

VI. EXPERIMENTS

This section outlines the experimental framework and discusses the results obtained when evaluating our dataset.[1]

Section VI-A first describes the experimental setup, including the DNN approaches, evaluation metrics, and key hyperparameters. Section VI-B then analyzes the performance of the proposed non-extended database, considering both delay-compensated and silence-suppressed variations. Section VI-C presents an ablation study to validate different design aspects. Finally, Sections VI-D and VI-E report the performance of the extended corpus.

A. Experimental Setup

The **experimental framework** has been designed to evaluate the **performance of RIRplay** under **out-of-domain** conditions (relative to that of ASVspoof 2019). In particular, we use **ASVspoof 2021** for evaluation, which stands out as one of the most realistic and challenging benchmarks for replay attack detection, but we also carry out an additional test on the ReMASC dataset to further support our conclusions.

We apply two different state-of-the-art (SOTA) well-known architectures for detection: a Conformer-based system [27] and AASIST [39], both using wav2vec 2.0 [40] for feature extraction. Other baseline DNN architectures, such as

[1]The full RIRplay corpora are publicly available on Zenodo: https://zenodo.org/records/19482955.

TABLE VII

MEAN AND STANDARD DEVIATION OF EER (%) (MEAN ± STD), WITH SUBTABLES (A) AND (B) CORRESPONDING TO CONFORMER AND AASIST. 'DC', 'SVAD' AND 'RVAD' DENOTE DELAY-COMPENSATION, SILEROVAD AND RVADFAST. LOWEST MEAN EER IN EACH COLUMN IS IN BOLD

Train set			Eval set				
			ASV2019		ASV2021		
	DC	sVAD	–	sVAD	–	sVAD	rVAD
ASV2019	–	×	**6.29 ± 0.78**	32.44 ± 4.74	40.26 ± 1.21	44.53 ± 2.75	42.83 ± 2.28
	–	✓	21.94 ± 2.61	**15.82 ± 1.23**	39.40 ± 1.19	36.89 ± 1.33	38.46 ± 0.76
RIRplay Non-Ext.	×	×	35.98 ± 0.79	42.64 ± 3.43	41.51 ± 1.83	41.91 ± 2.02	42.08 ± 1.12
	×	✓	38.41 ± 1.06	38.49 ± 1.09	36.51 ± 1.20	33.68 ± 0.72	35.43 ± 1.18
	✓	×	32.49 ± 0.97	31.56 ± 0.29	**33.72 ± 1.01**	34.71 ± 1.07	34.23 ± 0.84
	✓	✓	33.06 ± 0.39	31.94 ± 0.51	34.24 ± 1.31	**31.35 ± 1.22**	**32.78 ± 1.53**

(a) W2V2 + Conformer model.

Train set			Eval set				
			ASV2019		ASV2021		
	DC	sVAD	–	sVAD	–	sVAD	rVAD
ASV2019	–	×	**12.82 ± 2.08**	36.14 ± 2.85	42.24 ± 1.66	46.31 ± 1.88	45.35 ± 2.25
	–	✓	26.87 ± 2.35	**20.93 ± 1.11**	38.87 ± 1.02	37.91 ± 0.78	38.47 ± 0.78
RIRplay Non-Ext.	×	×	36.76 ± 3.77	39.47 ± 1.83	43.79 ± 5.58	44.91 ± 4.90	44.90 ± 5.70
	×	✓	39.76 ± 1.37	38.26 ± 0.32	34.52 ± 1.67	33.57 ± 0.41	33.95 ± 0.60
	✓	×	32.72 ± 1.00	33.78 ± 1.40	**34.50 ± 0.84**	34.53 ± 0.67	35.11 ± 0.99
	✓	✓	36.16 ± 3.09	32.06 ± 0.56	34.72 ± 0.83	**32.20 ± 0.56**	**33.46 ± 0.48**

(b) W2V2 + AASIST model.

LCNN [41] and RawNet2 [42], are not considered due to their comparatively poor performance in the ASVspoof 2021 Challenge [43] when trained on ASVspoof 2019, achieving EERs of 44.77% and 48.60%, respectively.

As evaluation metric, we use the Equal Error Rate (EER), reporting its average over 5 training runs per experiment (except for the ablation trials where only 3 runs are considered instead). The standard Adam optimizer [44] is employed with a batch size of 20, weight decay $3 \cdot 10^{-4}$, and a learning rate of 10^{-6} for Conformer and 10^{-7} for AASIST. Both architectures are trained for a maximum of 100 epochs, with an early stopping of 7 on validation. When considering VAD models, we use a threshold of 0.5 to detect non-speech segments.

Finally, it should be noted that all experiments were conducted on a Nvidia Tesla V100 GPU with 32 GB of VRAM, with each run taking approximately 8 hours.

B. Performance Comparison

Table VII (a and b) reports the average EER and standard deviation for both architectures, trained on different non-extended databases and evaluated on ASVspoof 2019 and 2021. The results demonstrate that system performance is highly dependent on the training data, as well as the effects of silence suppression and delay compensation, as listed below:

1) Silence suppression using sileroVAD (sVAD) reveals an inherent bias in ASVspoof 2019, leading to a significant performance drop when training on silence-suppressed data and testing on the original 2019 data, as evidenced by the results in the first two rows and first two columns of Tables VIIa and VIIb.
2) For our dataset, both architectures achieve lower EERs when trained on the delay-compensated (DC) version and evaluated on ASVspoof 2021, demonstrating the positive impact of delay compensation, as shown in the last three columns of results in both tables.
3) Additionally, applying VAD models to both the training and evaluation sets further improves performance by emphasizing voice-related features in both classes, particularly when using higher-quality training and evaluation data, such as RIRplay and ASVspoof 2021.
4) Training on RIRplay instead of ASVspoof 2019 leads to a notable EER reduction on ASVspoof 2021. Using the delay-compensated version with sVAD, the Conformer model achieves 31.35% EER vs. 36.89% when trained on ASVspoof 2019. A similar trend is observed for AASIST, with EER dropping from 37.91% to 32.20%. This result supports the use of more realistic acoustics adopted for RIRplay.

Drawing on the findings with sVAD, we observe that using a different VAD model for evaluation, such as rVADfast, produces almost identical results: 31.95% vs. 32.78% for Conformer, and 32.20% vs. 33.46% for AASIST. This suggests that performance is not highly sensitive to the specific VAD, emphasizing the importance of silence suppression for PA.

TABLE VIII

MEAN AND STANDARD DEVIATION OF THE EER (%) FOR DIFFERENT NOISE CONFIGURATIONS. ALL TRAINING SETS CORRESPOND TO RIRPLAY VARIANTS, WHERE 'CLEAN' DENOTES THE VERSION WITHOUT NOISE. THE LOWEST MEAN EER PER COLUMN IS HIGHLIGHTED IN BOLD

Train set	Eval set					
	RIRplay (training noises)			RIRplay (MUSAN noises)		ASV2021
	Clean	[25, 35] dB	[15, 25] dB	[25, 35] dB	[15, 25] dB	
Clean	$\mathbf{10.93 \pm 0.68}$	11.94 ± 0.56	11.95 ± 0.46	10.72 ± 0.61	12.02 ± 0.32	30.77 ± 0.87
Noise at [25, 35] dB	11.53 ± 0.38	$\mathbf{10.95 \pm 0.65}$	$\mathbf{10.67 \pm 0.84}$	$\mathbf{9.91 \pm 0.23}$	$\mathbf{10.09 \pm 0.41}$	$\mathbf{30.71 \pm 0.98}$
Noise at [15, 25] dB	12.90 ± 0.63	12.45 ± 0.72	10.82 ± 1.91	10.63 ± 0.91	10.65 ± 1.03	32.88 ± 0.96

C. Experimental Validation of Key Design Aspects

Although the RIRplay database yields a lower EER than ASVspoof 2019 when evaluated on ASVspoof 2021 (31.35% vs. 36.89%), it is important to show that essential design aspects contribute to improving the model's generalization capabilities across out-of-domain scenarios. The analysis conducted here is twofold. First, we focus on validating the **acoustic design** through ablation experiments, including decisions such as environment diversity and modeling ASV presentations where both attacker and speaker access the system from the same room. Second, we examine the influence of noise on replay detection performance, analyzing how different noise levels and out-of-domain conditions affect model generalization and robustness.

In the first case, for each ablation experiment, a modified RIRplay version is created by removing a specific design aspect. Models are evaluated under both **matched** and **cross-condition** settings: in the matched condition, training and testing are performed on the same dataset variant, whereas in the cross-condition setting, models are trained on a modified variant and evaluated on the original dataset, and vice versa. Two different aspects of the acoustic design are examined:

- The relevance of modeling the worst-case condition, where attacker and talker access the ASV system in the same room, is assessed by generating a dataset in which ASV presentations occur in different rooms ($r_1 \neq r_2$), i.e., enforcing $\text{RIR}_b^{r_1} \neq \text{RIR}_s^{r_2}$ (as in Figure 2a).
- The effect of reducing acoustic diversity is analyzed by limiting the dataset to 27 of the 112 environments from The ACE Challenge. The distribution of acoustic environments is now 10 for training, 7 for validation and 10 for evaluation.

In the second case, the **influence of noise** is examined by training and evaluating a clean RIRplay version, where the noise component $v[n]$ in Figure 2b is omitted, as well as a new RIRplay variant with increased noise levels, considering SNR_{rec} and SNR_{rep} in the range [15, 25] dB instead of the original [25, 35] dB. Additionally, to further assess the effectiveness of the incorporated noise, evaluation is performed on alternative RIRplay eval subsets generated using MUSAN noise excerpts [45].

The experiments described above are completed with evaluations on ASVspoof 2021 (with sVAD applied) in order to assess out-of-domain generalization. Given the high computational cost of fine-tuning wav2vec 2.0, we only conduct ablation experiments using three training runs with the Conformer model and the best RIRplay configuration (31.35% EER). The results are reported in Tables VIII and IX.

TABLE IX

MEAN AND STANDARD DEVIATION OF THE EER (%) FOR DIFFERENT ACOUSTIC ABLATION CONFIGURATIONS, INCLUDING SAME OR DIFFERENT ROOMS AND ENVIRONMENT REDUCTION. 'LESS ENVS' DENOTES THE LESS DIVERSE RIRPLAY DATASET VARIANT. FOR EACH CONDITION, THE LOWEST MEAN EER PER COLUMN IS HIGHLIGHTED IN BOLD

Train set	Eval set		
	Different room	Same room	ASV2021
Different room	10.97 ± 0.30	11.95 ± 0.24	32.30 ± 0.18
Same room	$\mathbf{10.63 \pm 0.69}$	$\mathbf{10.95 \pm 0.65}$	$\mathbf{30.71 \pm 0.98}$
	Less envs	**All envs**	**ASV2021**
Less envs	$\mathbf{9.67 \pm 0.41}$	12.44 ± 0.15	33.43 ± 1.52
All envs	9.87 ± 0.14	$\mathbf{10.95 \pm 0.65}$	$\mathbf{30.71 \pm 0.98}$

In general, as evidenced in Table IX, excluding any of the acoustic design aspects during training degrades performance when evaluating on datasets that include them. This degradation is particularly evident on ASVspoof 2021, where reduced acoustic variability leads to a notable EER increase (33.43% vs. 30.71%). This aspect is also reinforced by the results of the next Section VI-D (extended RIRplay). Although the performance gains from including noise are moderate, Table VIII shows that training with noisy data tends to improve robustness under both in-domain and out-of-domain conditions (MUSAN and ASVspoof 2021). The selected noise range [25, 35] dB achieves the lowest EER in noisy in-domain tests and provides the best overall generalization, yielding the lowest EER on both MUSAN and ASVspoof 2021 (30.71%) test sets. In contrast, higher noise levels ([15, 25] dB) degrade out-of-domain performance, particularly on ASVspoof 2021 (32.88%). Since DNNs trained on clean data are typically less robust to noise [46], and additive noise augmentation is widely considered beneficial [23], these results make the inclusion of a small amount of noise a reasonable choice to improve cross-condition generalization.

D. Final Results With Extended RIRplay

Since delay compensation and silence suppression yield improved generalization, Table X reports a comparison of EER results obtained by training the Conformer and AASIST

TABLE X

MEAN AND STANDARD DEVIATION OF EER (%) (MEAN ± STD) FOR ASVSPOOF 2019 AND RIRPLAY DATABASE USING THE W2V2 + CONFORMER (A) AND W2V2 + AASIST (B) MODELS. 'EXT.' REFERS TO EXTENDED VERSION. THE LOWEST MEAN EER IN EACH COLUMN IS HIGHLIGHTED IN BOLD

Train set			Eval set					
			ASV2019		ASV2021		ReMASC	
	DC	sVAD	–	sVAD	–	sVAD	–	sVAD
ASV2019	–	×	**6.29 ± 0.78**	32.44 ± 4.74	40.26 ± 1.21	44.53 ± 2.75	52.43 ± 1.40	47.56 ± 2.24
	–	✓	21.94 ± 2.61	**15.82 ± 1.23**	39.40 ± 1.19	36.89 ± 1.33	35.81 ± 1.43	38.90 ± 0.80
RIRplay	✓	✓	33.06 ± 0.39	31.94 ± 0.51	34.24 ± 1.31	31.35 ± 1.22	35.31 ± 2.28	34.66 ± 1.14
RIRplay Ext.	✓	✓	32.28 ± 0.63	30.82 ± 0.26	**31.86 ± 0.56**	**28.04 ± 0.49**	**34.57 ± 2.64**	**33.93 ± 1.55**

(a) W2V2 + Conformer model.

Train set			Eval set					
			ASV2019		ASV2021		ReMASC	
	DC	sVAD	–	sVAD	–	sVAD	–	sVAD
ASV2019	–	×	**12.82 ± 2.08**	36.14 ± 2.85	42.24 ± 1.66	46.31 ± 1.88	51.09 ± 2.95	46.67 ± 3.39
	–	✓	26.87 ± 2.35	**20.93 ± 1.11**	38.87 ± 1.02	37.91 ± 0.78	39.38 ± 4.42	40.52 ± 1.88
RIRplay	✓	✓	36.16 ± 3.09	32.06 ± 0.56	34.72 ± 0.83	32.20 ± 0.56	37.19 ± 3.95	36.30 ± 1.29
RIRplay Ext.	✓	✓	33.28 ± 0.74	31.29 ± 0.86	**31.95 ± 1.20**	**29.69 ± 0.93**	**34.57 ± 2.98**	**33.92 ± 1.14**

(b) W2V2 + AASIST model.

architectures on ASVspoof 2019 and on the extended, delay-compensated and silence-suppressed RIRplay dataset. As shown in Table Xa, this RIRplay variant leads to substantial performance gains for the Conformer system, achieving a best EER of 28.04% on ASVspoof 2021. This corresponds to an absolute EER reduction of 8.85% compared to training on ASVspoof 2019, which yields an EER of 36.89%. Although AASIST performs slightly worse, the doubled-size RIRplay dataset still provides an absolute EER improvement of 8.22% over ASVspoof 2019 (29.69% vs. 37.91%).

Moreover, evaluation on ReMASC further supports our findings. The results confirm that applying VAD during training reduces DNN error rates when detecting replay attacks in out-of-domain data (see sVAD results on ASVspoof 2019 and RIRplay compared to those without it). They also highlight the benefits of the extended RIRplay dataset, achieving EERs of 33.93% for the Conformer and 33.92% for AASIST.

To assess statistical significance, we compute 95% Confidence Intervals (CIs) for the Conformer EER values, averaged over five runs. Each model's error rate is treated as a Gaussian-distributed variable, following the theoretical approximation in [41] and [47]:

$$\mathrm{EER}_i \sim \mathcal{N}\left(\mathrm{EER}_i, \frac{\mathrm{EER}_i(1-\mathrm{EER}_i)}{4} \cdot \left(\frac{1}{N_{\text{bn.}}} + \frac{1}{N_{\text{sp.}}}\right)\right), \quad (3)$$

where $N_{\text{bn.}}$ and $N_{\text{sp.}}$ denote the number of bona fide and spoofed samples used to compute EER_i.

Since the reported EER is averaged over five independent runs, it is crucial to consider the variability between them. Assuming the EER values follow distinct Gaussian models, system dispersion can be approximated as a mixture of these distributions. This allows for estimating final CIs by incorporating both individual model uncertainty and variability from averaging. Figure 6 shows the resulting 95% CIs for the Conformer results, with similar outcomes observed for the AASIST model. A small CI for these systems indicates low variability and high consistency, reinforcing model reliability. The plot also highlights the performance gains from applying a VAD model, as evidenced by the separation between CIs of systems with and without VAD.

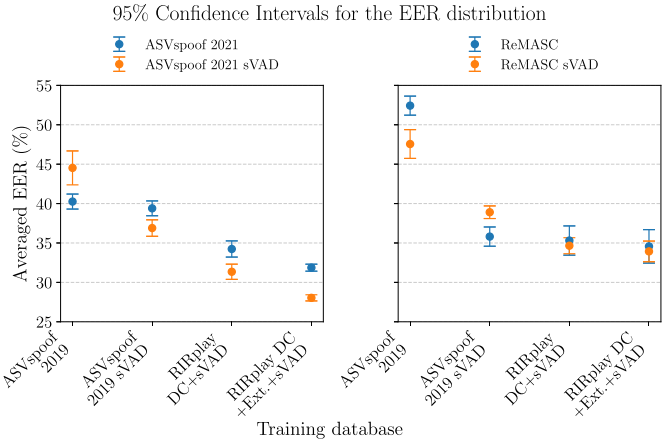


Fig. 6. 95% CIs for the EER distribution with Conformer architecture.

E. Further Analysis on ASVspoof 2021 Dataset

We performed a detailed breakdown of the performance across the loudspeaker, microphone, distance, and room

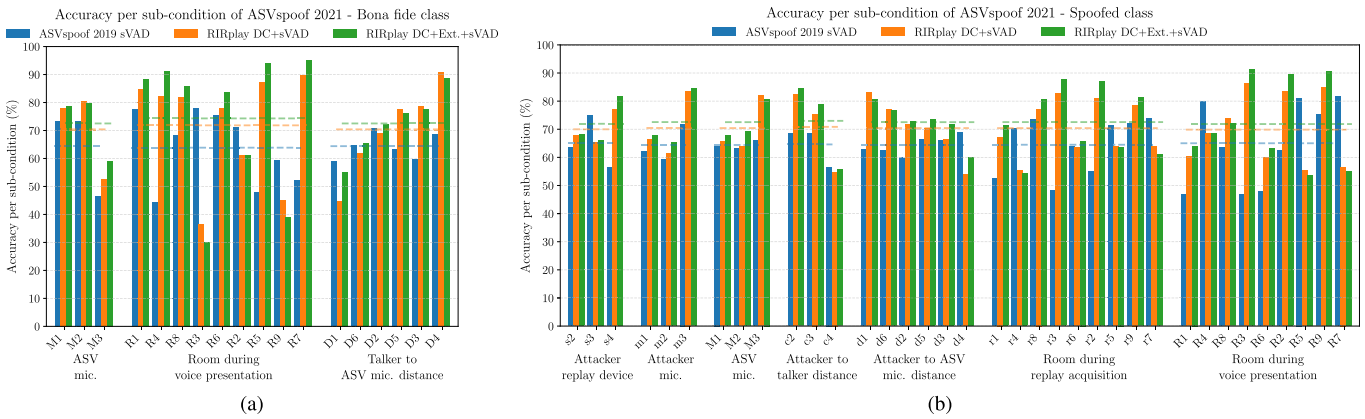


Fig. 7. Accuracy distribution for the bona fide (a) and spoofed (b) classes across all loudspeaker, microphone, distance, and room categories of ASVspoof 2021 [22], using different training sets and the best-performing Conformer run. Rooms and distances are ordered by volume and length, respectively, from largest to smallest. The horizontal dashed lines indicate the average accuracy within each condition category.

categories in the ASVspoof 2021 PA database. For this analysis, we select the best Conformer model among the five runs trained on the silence-suppressed versions of ASVspoof 2019, RIRplay, and extended RIRplay datasets. We computed the accuracy of each model per condition as follows:

$$\mathrm{Acc}(p) = \left(1 - \frac{N_{\mathrm{fails}}^{\mathrm{EER}}(p)}{N_{\mathrm{total}}(p)}\right) \times 100, \tag{4}$$

where $N_{\mathrm{fails}}^{\mathrm{EER}}(p)$ denotes the number of misclassified samples for condition p, determined using the EER decision threshold, and $N_{\mathrm{total}}(p)$, the total number of the evaluated samples for that ASVspoof 2021 condition.

In this way, we measure the proportion of correctly classified audios per condition using the EER-based threshold as the decision boundary, and present the results in Figure 7 as separate bar charts for the bona fide and spoofed classes. The different condition categories follow the official ASVspoof 2021 PA notation [22, Table III]. Categories R1-R9 (bona fide presentation) and r1-r9 (replay acquisition) denote nine different rooms of varying volume, from the largest R1/r1 ($8.0 \times 8.0 \times 2.4$ m^3) to the smallest R7/r7 ($4.5 \times 2.4 \times 2.4$ m^3). Distances D1-D6 (talker to ASV) and d1-d6 (attacker to ASV) range from 0.5 to 2.0 m, with D1/d1 and D6/d6 corresponding to the farthest positions and D4/d4 representing the closest and most centered configuration. For spoofed audios, s2-s4 indicate replay loudspeaker quality (low to high), while c2-c4 denote attacker-to-talker acquisition distances (1.5 m, 1.0 m, 0.5 m, respectively). As for the microphones, m1-m3 and M1-M3 (medium, high, and low quality) refer to the devices used during replay acquisition and ASV presentation, respectively.

As observed in Figure 7, performance across rooms does not reveal a clear trend for any training corpus, suggesting varying degrees of acoustic mismatch between training and evaluation data, consistent with the analysis of participant systems in the ASVspoof 2021 Challenge [22, Section III-B]. This underscores the need to increase training data diversity or apply additional processing strategies, such as curation or pruning [48]. Regarding the remaining conditions, Figure 7a shows that degraded acoustic scenarios, such as larger talker-to-ASV distances (e.g., D1 = 2.0 m) or low-quality ASV microphones (M3), tend to increase false rejection errors, as reflected by the lower bona fide accuracy. Conversely, Figure 7b exhibits the opposite trend: cleaner conditions, such as closer proximity (c4 and d4) or medium-to-high-quality microphones (m1/M1 and m2/M2), tend to increase false acceptance errors. These patterns appear more consistently in the RIRplay results, whereas ASVspoof 2019-trained models show more irregular behavior. Overall, the results confirm the robustness of the RIRplay-trained models, particularly the extended corpus, as evidenced by the higher average accuracies in Figure 7.

Finally, although our goal is not to directly compare our corpus results with the ASVspoof 2021 PA Challenge [43], using RIRplay instead of ASVspoof 2019 substantially improves the Conformer performance, achieving a competitive 28.04% EER on ASVspoof 2021, which would rank among the top five systems. Notably, this result is obtained with a single system modeling both bona fide and spoofed classes, unlike approaches that focus on a single class [49] or rely on multi-system fusion [50]. This represents a significant advancement, establishing a new baseline that we hope will boost further research on replay.

VII. Conclusion

In this work, we introduce RIRplay, a novel simulated dataset for physical access (PA) detection that faithfully replicates a real replay pipeline while overcoming key limitations of existing replay corpora. RIRplay provides more realistic and diverse room acoustics, a wider range of replay conditions, and realistic non-linear device modeling, making it a highly valuable resource for training DNN-based anti-spoofing systems.

The benefits of RIRplay are clearly demonstrated in our experiments. On the challenging ASVspoof 2021 benchmark, models trained on RIRplay achieve a substantial reduction in Equal Error Rate (EER), with the Conformer model reaching **28.04%** versus **36.89%** when trained on ASVspoof 2019, and AASIST dropping to **29.69%** from **37.91%**. These results represent a substantial advancement over those achieved with other corpora and confirm the suitability of the adopted design decisions.

Out-of-domain evaluation on the ReMASC dataset further reinforces these findings. Although absolute EERs are slightly higher (**33.93%** for Conformer, **33.92%** for AASIST) due to differing acoustic conditions, the relative performance remains

consistent, demonstrating that RIRplay-trained models maintain strong generalization.

Importantly, this work addresses the data insufficiency problem in replay detection rather than proposing a new CM architecture. RIRplay equips the research community with a rich, scalable, and realistic resource, enabling the development of more robust anti-spoofing systems capable of handling diverse environments, speakers, and playback devices.

Future directions include curating and pruning the dataset to address residual challenges in specific ASVspoof 2021 conditions; extending data diversity with more realistic acoustic environments; incorporating explicit models of attacker and ASV microphones; extending transfer learning to replayed deepfake detection; and exploring novel countermeasure systems that leverage RIRplay's stereo pairing to further boost robustness and cross-domain generalization.

ACKNOWLEDGMENT

The authors would like to thank the consortium responsible for the ASVspoof 2019 PA database development for providing them with the HHFR loudspeaker models employed in this work.

REFERENCES

[1] L. M. Mayron, Y. Hausawi, and G. S. Bahr, "Secure, usable biometric authentication systems," in *Proc. UAHCI*, 2013, pp. 195–204.

[2] A. Saleema and S. Thampi, *Voice Biometrics: The Promising Future of Authentication in the Internet of Things*. Hershey, PA, USA: IGI Global, 2018, pp. 360–389.

[3] A. E. Rosenberg, "Automatic speaker verification: A review," *Proc. IEEE*, vol. 64, no. 4, pp. 475–487, 1976.

[4] A. Mittal and M. Dua, "Automatic speaker verification systems and spoof detection techniques: Review and analysis," *Int. J. Speech Technol.*, vol. 25, no. 1, pp. 105–134, Mar. 2022.

[5] Y. Tu, W. Lin, and M.-W. Mak, "A survey on text-dependent and text-independent speaker verification," *IEEE Access*, vol. 10, pp. 99038–99049, 2022.

[6] O. Kohler and M. Imtiaz, "Investigation of text-independent speaker verification by support vector machine-based machine learning approaches," *Electronics*, vol. 14, no. 5, p. 963, 2025.

[7] H. Delgado, G. Ramondetti, E. Dalmasso, G. Karvitsky, D. Colibro, and H. Talib, "On deepfake voice detection—It's all in the presentation," 2025, *arXiv:2509.26471*.

[8] L. Li, Y. Chen, D. Wang, and T. F. Zheng, "A study on replay attack and anti-spoofing for automatic speaker verification," in *Proc. Interspeech*, Aug. 2017, pp. 92–96.

[9] H. A. Patil and M. R. Kamble, "A survey on replay attack detection for automatic speaker verification (ASV) system," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2018, pp. 1047–1053.

[10] M. Singh and D. Pati, "Countermeasures to replay attacks: A review," *IETE Tech. Rev.*, vol. 37, no. 6, pp. 599–614, Nov. 2020.

[11] M. Chica, A. Gomez-Alanis, E. Rosello, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "Database dependence comparison in detection of physical access voice spoofing attacks," in *Proc. IberSPEECH*, Nov. 2022, pp. 201–205.

[12] X. Wang et al., "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Comput. Speech Lang.*, vol. 64, Nov. 2020, Art. no. 101114.

[13] M. Todisco et al., "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 1008–1012.

[14] W. Mikulski, "Method of determining the sound absorbing coefficient of materials within the frequency range of 5 000–50 000 Hz in a test chamber of a volume of about 2 m^3," *Arch. Acoust.*, vol. 38, no. 2, pp. 177–183, Jun. 2013.

[15] B. Chettri, D. Stoller, V. Morfi, M. A. M. Ramírez, E. Benetos, and B. L. Sturm, "Ensemble models for spoofing detection in automatic speaker verification," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2019, pp. 1018–1022.

[16] Y. Gong, J. Yang, J. Huber, M. MacKnight, and C. Poellabauer, "ReMASC: Realistic replay attack corpus for voice controlled systems," in *Proc. Interspeech*, Sep. 2019, pp. 2355–2359.

[17] T. Kinnunen et al., "RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5395–5399.

[18] I. Yakovlev et al., "LRPD: Large replay parallel dataset," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6612–6616.

[19] R. Baumann, K. M. Malik, A. Javed, A. Ball, B. Kujawa, and H. Malik, "Voice spoofing detection corpus for single and multi-order audio replays," *Comput. Speech Lang.*, vol. 65, Jan. 2021, Art. no. 101132.

[20] Z. Wu et al., "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 4, pp. 588–604, Jun. 2017.

[21] T. Kinnunen et al., "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 2–6.

[22] X. Liu et al., "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2507–2522, 2023.

[23] J. C. Sanchez, A. M. Peinado, and A. M. Gomez, "Data augmentation techniques for physical access in voice anti-spoofing," in *Proc. IberSPEECH*, Nov. 2024, pp. 1–5.

[24] B. Chettri, E. Benetos, and B. L. T. Sturm, "Dataset artefacts in anti-spoofing systems: A case study on the ASVspoof 2017 benchmark," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 3018–3028, 2020.

[25] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6382–6386.

[26] A. Cohen, I. Rimon, E. Aflalo, and H. H. Permuter, "A study on data augmentation in voice anti-spoofing," *Speech Commun.*, vol. 141, pp. 56–67, Jun. 2022.

[27] 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*, Aug. 2023, pp. 5281–5285.

[28] C. V. Junichi Yamagishi and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," Centre Speech Technol. Res. (CSTR), Univ. Edinburgh, Edinburgh, U.K., Tech. Rep., Nov. 2019.

[29] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE challenge—Corpus description and performance evaluation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2015, pp. 1–5.

[30] E. Vincent. (2008). *Roomsimove*. [Online]. Available: http://homepages.loria.fr/evincent/software/Roomsimove1.4.zip

[31] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[32] J. M. Martín-Doñas, A. M. Peinado, I. López-Espejo, and A. Gomez, "Dual-channel speech enhancement based on extended Kalman filter relative transfer function estimation," *Appl. Sci.*, vol. 9, no. 12, p. 2520, Jun. 2019.

[33] R. Háber and L. Keviczky, "Nonlinear System Identification: Input-output modeling approach," in *Mathematical Modelling: Theory and Applications*. Norwell, MA, USA: Kluwer, Jul. 1999, p. 800.

[34] A. Novák, L. Simon, F. Kadlec, and P. Lotton, "Nonlinear system identification using exponential swept-sine signal," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 8, pp. 2220–2229, Aug. 2010.

[35] Y. Zhang, Z. Li, J. Lu, H. Hua, W. Wang, and P. Zhang, "The impact of silence on speech anti-spoofing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 3374–3389, 2023.

[36] B. Chettri and B. L. Sturm, "A deeper look at Gaussian mixture model based anti-spoofing systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5159–5163.

[37] Silero Team. *Silero VAD: Pre-trained Enterprise-Grade Voice Activity Detector (VAD), Number Detector and Language Classifier*. [Online]. Available: https://github.com/snakers4/silero-vad

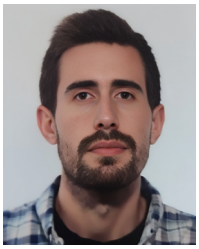
[38] Z.-H. Tan, A. K. Sarkar, and N. Dehak, "RVAD: An unsupervised segment-based robust voice activity detection method," *Comput. Speech Lang.*, vol. 59, pp. 1–21, Jan. 2020.

[39] J.-W. Jung et al., "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6367–6371.

[40] A. Babu et al., "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech*, Sep. 2022, pp. 2278–2282.
[41] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in *Proc. Interspeech*, Aug. 2021, pp. 4259–4263.
[42] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6369–6373.
[43] J. Yamagishi et al., "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in *Proc. Ed. Autom. Speaker Verification Spoofing Countermeasures Challenge*, Sep. 2021, pp. 47–54.
[44] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[45] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," 2015, *arXiv:1510.08484*.
[46] A. Pervaiz et al., "Incorporating noise robustness in speech command recognition by noise augmentation of training data," *Sensors*, vol. 20, no. 8, p. 2326, Apr. 2020.
[47] S. Bengio and J. Mariéthoz, "A statistical significance test for person authentication," in *Proc. Speaker Odyssey*, 2004, pp. 237–244.
[48] D. Combei, A. Stan, D. Oneata, N. Müller, and H. Cucu, "Unmasking real-world audio deepfakes: A data-centric approach," in *Proc. InterSpeech*, 2025, pp. 5343–5347.
[49] X. Wang, X. Qin, T. Zhu, C. Wang, S. Zhang, and M. Li, "The DKU-CMRI system for the ASVspoof 2021 challenge: Vocoder based replay channel response estimation," in *Proc. ASVspoof Challenge Workshop*, 2021, pp. 16–21.
[50] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva, "STC antispoofing systems for the ASVspoof2021 challenge," in *Proc. Ed. Autom. Speaker Verification Spoofing Countermeasures Challenge*, Sep. 2021, pp. 61–67.

Jose C. Sanchez-Valera received the B.Sc. and M.Sc. degrees in telecommunication engineering from the University of Granada, Spain, in 2021 and 2023, respectively, where he is currently pursuing the Ph.D. degree in information and communication technologies. Since 2023, he has been a Ph.D. Researcher with the Department of Signal Theory, Telematics and Communications, University of Granada, where he is involved in research on speech processing, voice biometrics, anti-spoofing systems, and replay detection for secure voice interaction.

Antonio M. Peinado (Senior Member, IEEE) received the M.S. and Ph.D. degrees in electronic physics from the University of Granada, Spain, in 1987 and 1994, respectively. Since 1988, he has been with the University of Granada, where he has led several research projects related to speech/image processing and transmission. In 1989, he was a Consultant with the Speech Research Department, AT&T Bell Labs, Murray Hill, NJ, USA. In 2018, he was a Visiting Scholar with the Language Technologies Institute, Carnegie Mellon University (CMU), Pittsburgh, PA, USA. He is currently a Full Professor with the Department of Signal Theory, Networking and Communications, University of Granada, and the Head of the Signal Processing, Multimedia Transmission, and Speech/Audio Technologies (SigMAT) Research Group. He has authored numerous publications in international journals and conferences, and has co-authored the book *Speech Recognition Over Digital Channels: Robustness and Standards* (Wiley, 2006). His current research interests focus on various speech technologies, including antispoofing, speech enhancement, and robust speech recognition and transmission. He has served as a reviewer for numerous international journals and conferences, an evaluator for project and grant proposals, and a member of the technical program committee for several international conferences.



Juan M. Martin-Doñas received the M.Sc. degree in telecommunication engineering and the Ph.D. degree in information and communication technologies from the University of Granada, Spain, in 2017 and 2021, respectively. From 2016 to 2020, he was a Ph.D. Fellow with the Department of Signal Theory, Telematics and Communications, University of Granada. From 2020 to 2024, he was a Research Scientist with the Department of Speech and Natural Language Technologies, Fundación Vicomtech, Spain. He is currently an Assistant Professor with the Department of Industrial Engineering, University of La Laguna, Tenerife, Spain. His main research interests include digital speech processing, anti-spoofing countermeasures, and audio deepfake detection.

Alejandro Gomez-Alanis received the M.Sc. degree in telecommunications engineering and the Ph.D. degree in artificial intelligence from the University of Granada, Spain, in 2017 and 2022, respectively. In 2016 and 2017, he was a Software Engineer at Celtiberian Solutions, where he developed automatic statistical tools and mobile/web applications. From 2017 to 2022, he held an FPU Fellowship with the Department of Signal Theory, Telematics and Communications, University of Granada, where he conducted research in speech biometrics. He subsequently worked as an Applied Scientist at Amazon Alexa, a Machine Learning Engineer at Procore Technologies, and a Research Scientist at Meta. His work has focused on speech technologies, biometrics, computer vision, ad ranking, and deep learning. His research interests include speech processing, speech modelling, and applied artificial intelligence for human-centered applications.

Angel M. Gomez received the M.Sc. and Ph.D. degrees in computer science from the University of Granada, Spain, in 2001 and 2006, respectively. In 2002, he joined the Department of Signal Theory, Telematics and Communications, University of Granada, where he is currently a Full Professor. He is also a member of the Signal Processing, Multimedia Transmission and Speech/Audio Technologies (SigMAT) Research Group. His research interests include speech processing, audio technologies, and machine learning.

Massimiliano Todisco (Member, IEEE) is currently an Associate Professor with the Digital Security Department, EURECOM, Sophia Antipolis, France. His work explores the design of robust and trustworthy machine learning systems through approaches grounded in generative artificial intelligence, adversarial learning, and privacy-preserving techniques, including watermarking. He is also interested in improving the interpretability and reliability of biometric systems in real-world applications. He is best known for proposing the constant Q cepstral coefficients (CQCCs), which contributed to advances in fake audio detection and earned the ISCA Best Paper Award from 2015 to 2019. His research interests include voice biometrics and related machine learning techniques for reliable and trustworthy systems. He serves as an Associate Editor for IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE. He is a Co-Organizer of international challenges, such as ASVspoof and VoicePrivacy.