Computer Speech & Language 65 (2021) 101132

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Voice spoofing detection corpus for single and multi-order audio replays☆

Roland Baumann[a], Khalid Mahmood Malik*,[a], Ali Javed[a], Andersen Ball[a], Brandon Kujawa[a], Hafiz Malik[b]

[a] Computer Science and Engineering Department, Oakland University, Rochester, MI, USA
[b] Electrical and Computer Engineering, University of Michigan-Dearborn, MI, USA

ARTICLE INFO

Article History:
Received 26 August 2019
Revised 16 June 2020
Accepted 9 July 2020
Available online 16 July 2020

Keywords:
Multi-order voice replay attack
Internet of multimedia things
Voice replay spoofing
Voice controlled devices
Automatic speaker verification anti-spoofing
Voice spoofing dataset

ABSTRACT

The evolution of modern voice-controlled devices (VCDs) has revolutionized the Internet of Things (IoT) and resulted in the increased realization of smart homes, personalization, and home automation through voice commands. These VCDs can be exploited in IoT driven environments to generate various spoofing attacks, including the chaining of replay attacks (i.e. multi-order replay attacks). Existing datasets like ASVspoof 2017, ASVspoof 2019, and ReMASC contain only first-order replay recordings (i.e. replayed once); therefore, they cannot offer evaluation of anti-spoofing algorithms capable of detecting multi-order replay attacks. Additionally, large-scale datasets like ASVspoof 2017 and ASVspoof 2019 do not capture the characteristics of microphone arrays, which are an essential characteristic of modern VCDs. Therefore, there exists a need for a diverse replay spoofing detection corpus that consists of multi-order replay recordings against bona fide voice samples. This paper presents a novel voice spoofing detection corpus (VSDC) to evaluate the performance of multi-order replay anti-spoofing methods. The proposed VSDC consists of first-order (i.e. replayed once) and second-order replay (i.e. replayed twice) samples against the bona fide audio recordings. We ensured to create a diverse replay spoofing detection corpus in terms of environments, recording and playback devices, speakers, configurations, replay scenarios, etc. More specifically, we used 35 microphones, 25 different recording configurations, and 60 different playback configurations for first- and second-order replays to generate a total of 14,050 samples belonging to 19 speakers. Additionally, the proposed VSDC can also be used to evaluate the performance of speaker verification systems in terms of independent speaker verification. To the best of our knowledge, this is the first publicly available replay spoofing detection corpus comprised of first and second-order replay samples. Experimental results signify the effectiveness of the proposed VSDC in terms of evaluating the performance of anti-spoofing methods under multi-order replay attacks and diverse conditions.

1. Introduction

The growing trend of personalization, the realization of smart homes, and the desire for easy control of home devices are driving factors for the tremendous evolution of Internet of Things (IoT) devices. Voice assistants, the user interface of voice-controlled

☆ This paper is submitted on 26 August, 2019 and resubmitted on February 10, 2020.
*Corresponding author.
E-mail address: mahmood@oakland.edu (K.M. Malik).

https://doi.org/10.1016/j.csl.2020.101132
0885-2308/

devices (VCDs) such as Google Home, Amazon Alexa, and Apple Siri are becoming an essential component of IoT. VCDs are designed around audio and video multimedia capabilities, they make popular a new sub-field of IoT, called the Internet of Multimedia Things (IoMT) (Alvi et al., 2015). IoMT devices, a subset of IoT devices, are equipped with microphones, cameras, and speakers. Likewise, many connected toys (children, 2020) with voice interfaces are becoming part of the Internet of Toys (IoToys) (Chaudron et al., 2019) and hence IoT. VCDs are susceptible to various audio spoofing attacks such as replay attacks, voice cloning attacks, and laser-based audio injection attacks (Sugawara et al., 2019) etc., while IoMT devices face various multimedia spoofing challenges including deepfakes (Agarwal et al., 2019).

Voice assistants have enabled enormous connectivity among VCDs, opening new vistas of research (Malik et al., 2019). Notably, the addition of microphone arrays and speakers allow these devices to engage in two-way communication, allowing them to play audio and accept voice commands from other IoT devices. The most recognizable feature of VCDs has been the capability to connect all household IoT devices together with voice commands. Voice assistants are now being directly integrated into thermostats, refrigerators, light switches, entertainment systems, and cars. It is essential to mention that many IoT devices in smart homes are controlled remotely through VCDs. In addition to controlling IoT devices in the home, integrated voice assistants are also being used in a variety of Internet based applications related to VCDs such as entertainment, communication, shopping, healthcare, business, banking services, etc. With so many devices in a home being able to provide a voice assistant, the system resembles the dream of science fiction television, where an omnipresent computer is continuously ready to provide quick and easy verbal access. VCDs themselves could be used to replay audio to each other forming the basis of multi-hop scenarios. The open space inside a home becomes a transmission medium through which one VCD can replay voice commands to another VCD. Fig. 1 illustrates how a smart home can have many VCDs capable of speaking to each other.

Most VCDs are equipped with array microphones, which means they have more than one microphone. The Amazon Echo Dot 3 uses an array of 4 microphones. This array of microphones allows the VCD to determine the location of the speaker, selection of the best microphone and use the other microphones to reject background noise. This configuration enables VCDs to pick up voice commands at long distances (few meters) in less than ideal conditions. This fact enables the VCD to be more susceptible to replay attacks.

Automated Speaker Verification (ASV) systems have advanced in recent years and their application are increasing in a variety of real-world authentication scenarios involving both logical and physical access (Sahidullah et al., 2019). The applications of ASV are expected to be more ubiquitous in the future due to pervasiveness of smart speakers, smartphones, and other voice-enabled smart devices. Audio-specific spoofing attacks (Sahidullah et al., 2019) on ASV can be categorized into replay, speech-synthesis (SS), voice conversion (VC), and impersonation. Among all audio spoofing attacks, replay attacks could be more prevalent in the future, as less tech-savvy intruders can generate them to disrupt the ASV system of a VCD (Todisco et al., 2017). Existing spoofing datasets (ASVspoof 2017 dataset, 2019a; ASVspoof 2019 dataset, 2019b; Gong et al., 2019; RedDots Project, 2020) are designed for evaluation of testbeds that consider replay spoofing as a binary-class problem. These datasets have been used in addressing the scenario of a one-time replay, in applications such as voice-driven banking. However, we have demonstrated through

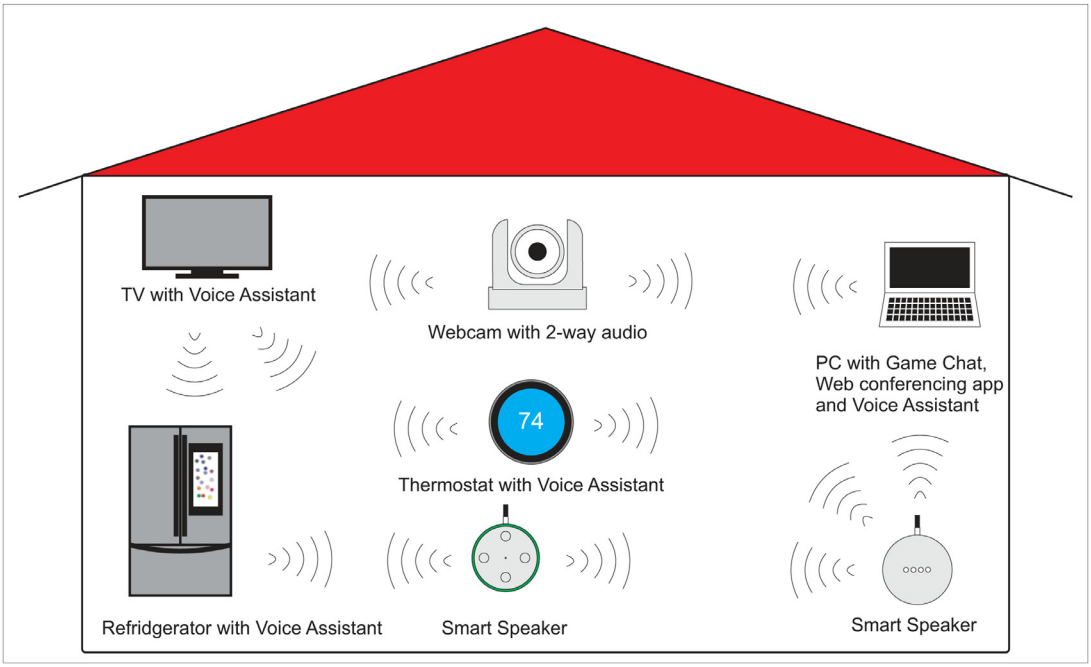

Fig. 1. VCD connectivity in the home for sending/receiving audio.

experimentation in our earlier work (Malik et al., 2019) that VCDs are very vulnerable to second-order replay attacks and are unable to classify between the original and spoof samples in multi-hop scenarios. Second-order replays can contain significant noise due to each subsequent replay contributing noise. Capturing replayed audio with this additional noise can be beneficial in ASV research. Research has shown that a limitation of early ASV methodology has been using clean speech corpus prepared in laboratory environments (Sahidullah et al., 2019).

This vulnerability of VCDs to multi-order replays can easily be exposed by an intruder to cause severe financial loss and data theft. Additionally, existing datasets (e.g. ASVspoof 2017 and ASVspoof 2019) do not contain the audio samples recorded from devices having array microphones, particularly in chained replay scenarios, which are quite common in IoMT. Therefore, there exists a need to create a replay spoofing dataset to evaluate applications and testbeds that may involve multi-hop voice propagation scenarios and samples recorded with devices having microphone arrays. For this purpose, we designed a novel Voice Spoofing Detection Corpus (VSDC). We developed this large-scale dataset to serve as an audio forensic testbed. Our dataset is the first dataset to contain multi-order replays consisting of bona fide, first- and second-order replay samples that can effectively be used to evaluate the performance of anti-spoofing methods in multi-hop scenarios. We ensured that we created a diverse replay spoofing detection corpus in terms of environment, recording and playback devices, speakers, configurations, replay scenarios, etc. More specifically, we used 35 microphones, 25 unique recording configurations, 60 unique playback configurations to generate a total of 14,050 samples belonging to 19 speakers of different ages and genders. Unlike traditional ASV systems which consider replay detection as a binary class problem, chained VCDs consider it as a multiclass problem. Because it is possible for a certain VCD, which itself has robust binary spoof countermeasure, to receive played back voice from other VCDs that are either compromised or prone to voice spoofing attacks due to a weak or absent spoof countermeasure (Malik et al., 2020).

All four datasets ASVspoof 2019, ASVspoof 2017, ReMASC and VSDC are designed with assisting in the development of countermeasures to voice replay attacks. The ReMASC and VSDC datasets specifically targeted and studied voice replay attacks on VCDs. These datasets could be used to develop and evaluate the countermeasures against voice replay attacks on VCDs. Also, these datasets could be used to develop anti-spoofing solutions for other voice-driven systems. The ASVspoof 2017 dataset uses a subset of the RedDots (RedDots Project, 2020; Kinnunen et al., 2017) dataset for its bona fide recordings. The ASVspoof 2019 dataset uses the VCTK (Corpus, 2020) database for a part of its set. In ASVspoof 2019 replays were created by simulation; this was done by prepossessing bona fide speech from the VCTK database.

Tables 1 and 2 highlight the differences between VSDC, ReMASC, and ASVspoof datasets (2017 and 2019). The salient features of VSDC are the inclusion of second-order replay samples, and use of array microphones along with non-array microphones (e.g. professional-grade microphones) at both first-order replay and second-order replay, which is needed to capture the characteristics of IoT applications. Table 2 further highlights the number of recording devices, recording microphones, playback devices, and environments used to develop these datasets. The ReMASC dataset used fewer microphones for recording bona fide speech as their emphasis was placed on using the array microphones found in VCDs. The focus for VSDC was to use a large variety of microphones to test how the audio signal of the replays changed between different microphones.

2. The landscape of multi-hop replay attacks

In this section, we briefly discuss the landscape of replay attacks involving VCDs. The addition of a voice interface introduces a new attack surface to be exploited in homes, offices, businesses, and hospitals. These scenarios demonstrate that multiple replays on VCDs can be used to exploit systems having voice interfaces. Although we discuss the scenarios of smart homes in this paper, threats associated with replay attacks can go beyond homes, as voice-controlled applications are being developed for smart cities, futuristic cars, and businesses. Amazon has already launched its smart assistant Alexa for business automation (First the Home, 2019). Currently, Amazon is working on healthcare apps that use smart speakers to perform various tasks (Amazon, 2019). Details of a few representative scenarios involving the multi-order replay attacks are discussed below. It is important to mention that we have experimentally verified these scenarios.

2.1. Scenario 1: Webcam replay

Shown in Fig. 2 are the two scenarios where VCDs are used to replay audio to each other. In scenario 1, shown in Fig. 2(a), a compromised webcam listens to a user giving commands to a Google Home device. In such a scenario, a webcam can be accessed by compromising the homes WiFi network using a tool (e.g. Aircrack-NG (Aircrack-Ng, 2019)). The study of Common

Table 1
Comparison of datasets in terms of numbers of speakers, genuine and replayed recordings (Gong et al., 2019; Delgado et al., 2018; Todisco et al., 2019).

Dataset	Speakers	Genuine	Replayed	Replayed (second-order)
ASVspoof 2019	107	12,960	116,640	NA
ASVspoof 2017	42	3,565	14,465	NA
ReMASC	55	9,240	45,472	NA
VSDC	19	1,687	6,179	6,184

Table 2
Comparison of datasets in terms of recording devices, recording microphones, playback devices and environments (Gong et al., 2019; Delgado et al., 2018; Todisco et al., 2019).

Dataset	Recording Devices	Recording Microphones	Playback Devices	Environments
ASVspoof 2019	40	40	40	27
ASVspoof 2017	25	25	26	22
ReMASC	2	2	4	4
VSDC	31	35	10	15

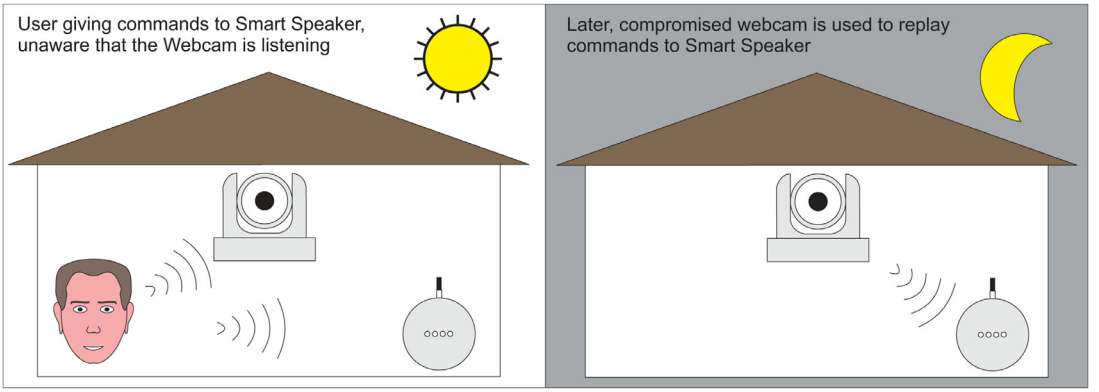

Fig. 2. Three replay attack scenarios. a) Webcam Replay. b) Drop-In Replay. c) Drop-In Multi-Home Replay.

Vulnerability and Exposures (CVEs) (CVE-2019-3948, 2019) shows that there exist many vulnerabilities that allow unauthenticated access to webcams these days. After capturing audio of the victim at home, the attacker can use the webcam to replay commands to a Google Home device in the absence of the victim. This demonstrates a traditional replay attack with only one point of replay. The webcam in this scenario could also be a baby monitor or other compromised VCD. We have verified through experiments that Google Home devices only authenticate the user based on the wake word "Hey Google". As long as the audio recording of the user saying, "Hey Google" in this scenario is clear enough, then an attacker can replay the wake phrase and insert any subsequent command (e.g. "Open Garage Door").

2.2. Scenario 2: Drop-In replay

In scenario 2, shown in Fig. 2(b), we describe a situation where an attacker obtains an original recording of a victim. In such a scenario a co-worker or someone close to the victim can use social engineering techniques to direct a normal conversation in a way that gets the victim to say the commands that the attacker is looking for. The attacker would record the entire conversation without the victim's knowledge and would be able to cut out the clips that he wants to use in the attack. The attacker intends to replay the audio from his home to the victim's home. The Amazon Echo devices offer a feature that allows audio conferencing between these devices. This feature, called Drop-In, works between different homes and Echo devices owned by different people if their contact list permissions are set to allow that contact to Drop-In. When using the Drop-In mode, the receiving Amazon Echo device plays a small chime and changes the devices light ring to green, thus enabling the conference mode. The presence of the recipient of the conference call is not required as the conference mode is enabled without any additional verification. If another VCD is nearby, then commands can be replayed through the audio conference.

For example, if an attacker is able to add himself, to Bob's contact list and allow himself the Drop-In permission, then the attacker could start the audio conference between Amazon Echo VCDs at any time. By simply asking their Amazon Echo to "Drop-in on Bob". The attacker could then replay the command "Hey Google, open the garage door". Although this scenario requires that Bob's Amazon Alexa contact list be exploited by the attacker gaining access to Bob's smartphone, this scenario is plausible, as a smartphone can be accessed by a trusted individual such as a friend, co-worker, or child.

This replay attack scenario demonstrates a proof-of-concept that VCDs in the home are vulnerable to replay attacks as long as the victim's audio can be played in front of a VCD in the home. This scenario is an example of a multi-hop replay attack as the original audio is replayed once to an Amazon Echo and then replayed again from the victim's Amazon Echo device. We observed during the multi-hop replay scenarios that signal degradation due to multiple replays are unable to cause any problem as long as the playback audio is audible. While this scenario used two Amazon Echo devices, they could be replaced with other devices capable of transmitting audio such as a smartphone app being used to send audio to a webcam in another home.

2.3. Scenario 3: Drop-In multi-Home replay

In scenario 3 shown in Fig. 2(c), we describe a situation where an attacker has remotely managed to exploit a vulnerability in a webcam to access its audio and video streams. The camera is conveniently located in the kitchen near an Amazon Echo device. The attacker can quickly identify the presence of the victim at home and observe their interactions with the VCD by accessing the camera. The Amazon Echo device itself does not have any voice verification system to verify the authenticity of any person. The attacker can cause chaos at the victim's home during his absence by issuing commands to change the thermostat settings, turning on and off the lights, opening the garage door, etc. from the webcam. Every IoT device in the home that is connected to the Amazon Echo is now accessible to the attacker.

The Amazon Echo Drop-In feature causes an additional threat in this situation. If the homeowner has allowed the Drop-In mode between friends, family members, or with work colleagues, then the attacker will also have access to start an audio conference between other Amazon Echo devices. At this point, the attacker could Drop-In to another Amazon Alexa located in another family members home. If there happens to be another VCD nearby, the attacker can then attempt to control IoT devices in the second home as well. In this scenario, the attacker can easily control another home remotely through the Drop-In feature.

While we proposed a few scenarios, it can be assumed that the device with the weakest security will be exploited. Some VCDs such as webcams and toys with voice driven interfaces (frontdoor, 2020) are known to have vulnerabilities that expose their credentials or audio streams due to the fast and inexpensive manner of their production. Once a VCD has been exploited, then an attacker can have multiple options from listening and collecting audio to replaying audio or cloned voices.

3. Dataset

This paper presents a unique voice spoofing detection corpus consisting of bona fide, first- and second-order replay recordings by setting up different scenarios of chained VCDs. This multi-hop replay feature in our corpus can be used to evaluate the performance of different replay anti-spoofing algorithms under diverse recording and playback environments, configurations, and devices. Our proposed VSDC can also be used to evaluate the performance of speaker verification systems as our corpus includes audio samples of 19 different speakers. Additionally, VSDC adds more diversity to existing datasets in terms of sampling rate. For example, the sampling rate of 96,000 HZ of VSDC, complements other datasets such as ASVspoof 2017, ReMasc, and ASVspoof 2019 which have sampling rates of 16,000 HZ and 44,100 HZ. The minimum length of each audio sample in our dataset is 6

seconds in duration. Note that for original recordings, we have used both professional-grade microphones and cell phones. In order to obtain data using cellphones/tablets, we designed and released Android and iPhone applications.

3.1. Definitions and data collection strategy

As this paper discusses the idea of multiple points/order of replays, we need to define the terminology used to specify the given point of replay. We will refer to bona fide recordings of a person giving voice commands to a VCD as the zero point of replay or (0PR). When the original recording is replayed from an audio speaker, we will refer to the output audio as the first point of replay or (1PR). Similarly, when the 1PR audio is replayed through a chained VCD, we refer to that output audio as the second point of replay or (2PR).

Shown in Fig. 3 is the process of capturing audio to create the dataset consisting of 0PR, 1PR, and 2PR. The bona fide phrases (0PR files) can be captured on any recording device. The 0PR files are then copied to a PC for generating the replays. The PC replays the 0PR file through an audio speaker creating the 1PR audio. VCDs are set up in an audio conference mode so that the audio played at 1PR is replayed by the VCD at 2PR. The PC used for creating the data sample is simultaneously replaying the 0PR file while capturing the resulting 1PR and 2PR audio. The USB sound card connected to the PC in Fig. 3 can be replaced by the onboard sound card of the PC or with a sound interface box.

For the data collection, we used the Audacity tool (Application, 2019) to simultaneously play the 0PR audio while capturing the resulting 1PR and 2PR audio. The Audacity tool can play from one audio track while simultaneously recording audio on other tracks. To capture replays, a scenario such as two Amazon Echo devices in conference mode (Drop-In) is set up to create a chain of VCDs. Audacity is set up with the bona fide (0PR) recording on track 1. The PC plays the audio through a connected audio speaker; this output audio is recorded on track 2 by Audacity and becomes the 1PR recording. At the same time, the VCD replays the audio to the next device in the chain. The resulting output is recorded on track 3 by Audacity and becomes the 2PR recording. Using this method, we captured the 1PR and 2PR replays at the same time. It was necessary to maintain proper isolation between the 1PR and 2PR environments to ensure that each respective microphone would not receive the same sound as the other (e.g. the 2PR microphone would not overhear the sound coming from the audio speaker used in the 1PR environment). Audacity is then used to trim the samples into 6-second lengths. All data samples at 0PR, 1PR, and 2PR are exported as separate files. The reasoning for the WAV files all being 6 seconds in length is to make the automatic clipping of the samples easier. Note that all 0PR are replayed in our settings to get 1PR and 2PR simultaneously.

Recording was performed at a sample rate of 96,000Hz using a 32-bit float format. The exception to this was when Bluetooth speakers were used for replays, a recording sample rate of 48,000Hz was instead used. This was due to the fact that the Bluetooth speakers only supported a playback rate of 48,000Hz and Audacity would only allow simultaneous recording and playback at the same rate. The resulting files were re-sampled to 96,000Hz. All files have been exported as WAV files at 96,000Hz 32-bit.

3.2. Voice commands and recording subjects

We used 42 different command phrases to create the bona fide recordings as shown in Table 3. All commands start with the activation phrase “Hey Google”, “Computer” or “Alexa”. Some of the voice commands used for recordings include, “Hey Google, turn on the kitchen light” and “Computer, turn off living room light”. The phrasing of a given command using the activation word “Computer” reflects that replay attacks are not a vendor-specific issue. A total of 19 speakers, ages 18–60 years old, participated in data collection. Out of 19 speakers, 10 are male, and 9 are female. Some of the speakers are not native English speakers. Each speaker recorded the original file by repeating a given set of phrases typical of commands given to VCDs. Some of the volunteers

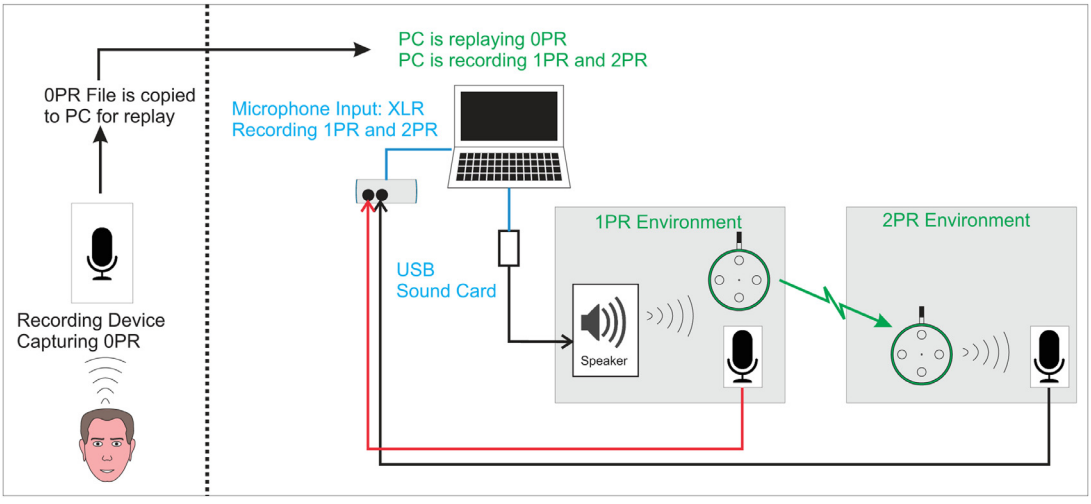

Fig. 3. Dataset creation configuration.

Table 3
Phrases used for 0PR samples.

Phrases	
Computer, turn on office lamp	Hey Google, turn off bedroom lamp
Computer, turn off office lamp	Hey Google, turn on living room light
Computer, turn on kitchen lights	Hey Google, turn off living room light
Computer, turn off kitchen lights	Hey Google, who am I
Computer, turn on bedroom lamp	Hey Google, give me an Easter egg
Computer, turn off bedroom lamp	Hey Google, good morning
Computer, turn on living room light	Hey Google, tell me a joke
Computer, turn off living room light	Hey Google, beam me up
Computer, who am I	Hey Google, set phasers to kill
Hey Google, turn on office lamp	Hey Google, tea, earl grey, hot
Hey Google, turn off office lamp	Hey Google, my name is Inigo Montoya
Hey Google, turn on kitchen lights	Hey Google, I want the truth
Hey Google, turn off kitchen lights	Hey Google, turn on bedroom lamp
Alexa, turn on office lamp	Alexa, my name is Inigo Montoya
Alexa, turn off office lamp	Alexa, I want the truth
Alexa, turn on kitchen lights	Alexa, turn off kitchen lights
Alexa, turn on bedroom lamp	Alexa, turn off bedroom lamp
Alexa, turn on living room light	Alexa, turn off living room light
Alexa, give me an Easter egg	Alexa, good morning
Alexa, tell me a joke	Alexa, beam me up
Alexa, set phasers to kill	Alexa, tea, earl grey, hot

recorded the original phrase sets multiple times using different microphones in diverse environments. In total, 173 different 0PR source sets were created, each set being a specific individual under specific recording conditions. A total of 1,687 0PR source phrases were spoken; the exact number of phrases each individual spoke varied, with each speaker recording no fewer than 9 phrases for a set.

3.3. 0PR Environments

The 0PR environment is the area where the bona fide sound samples are recorded. The VSDC includes samples recorded at 10 different unique environments (for 0PR) that contain different amounts of ambient noise (Fig. 4). We recorded the samples in different environments to ensure diversity. The environments considered "noisy", are the Computer Lab with Music, Car Off with light rain, Car on with Light Rain, Cafeteria, and University Court Yard. The Computer Lab with Music environment contained ambient noise from loud music. Both the Car on with Light Rain and Car Off with Light Rain environments are deemed as noisy due to the consistent pattering of rain and surrounding cars. The Cafeteria environment contains ambient noise from student activities. The University Courtyard environment contains rustling trees and background conversations. The environments classified as low noise are, the Office Desk, Kitchen Table, Bedroom, and Computer Lab. All indoor environments contain some background noise from air circulation systems.

3.4. Playback environments

For playback environments, we used a living room, home office desk, copy room, mini audio booths, conference room, lab classrooms to create the 1PR and 2PR replays. A brief description of each environment is given below.

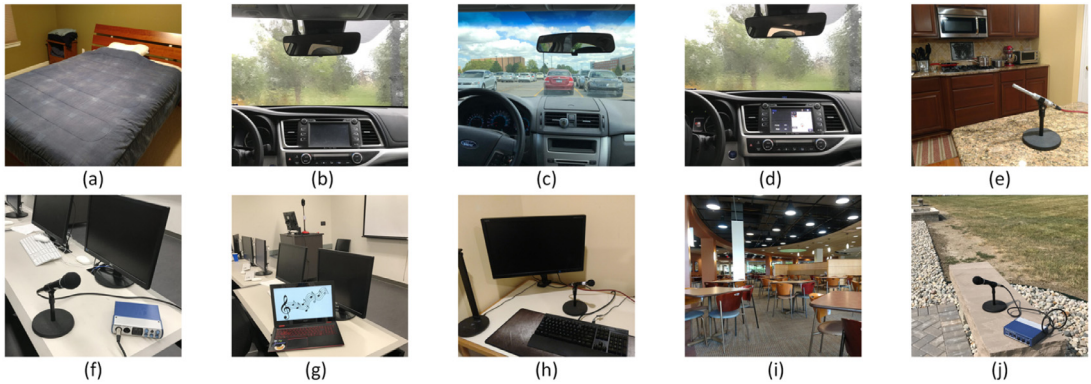

Fig. 4. Environments used for 0PR recordings. (a) Bedroom, (b) Car Off Light Rain, (c) Car On, (d) Car On Light Rain, (e) Kitchen Table, (f) Computer Lab, (g) Computer Lab with Music, (h) Office Desk, (i) University Cafe, (j) University Courtyard.

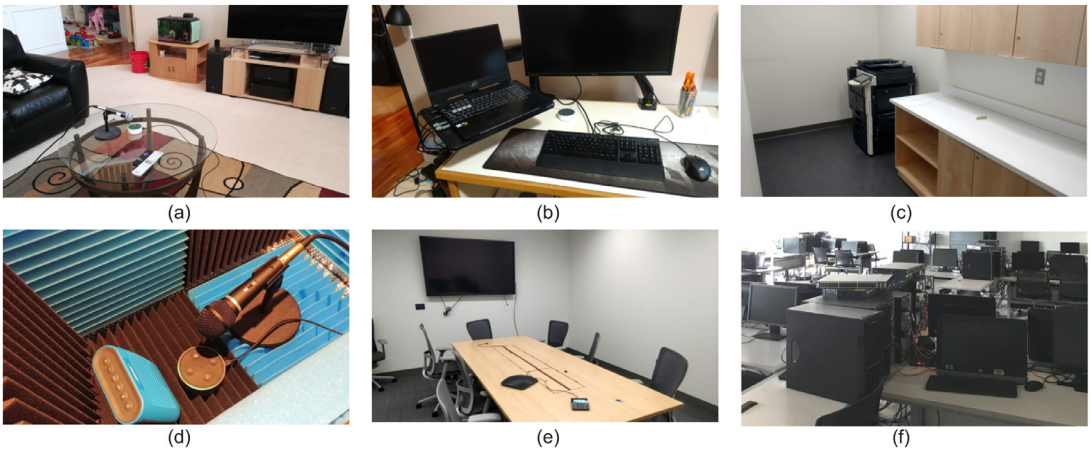

Fig. 5. Playback Environments. (a) Living Room, (b) Home Office, (c) Copy Room, (d) Inside Mini Recording Booth, (e) Conference Room, (f) Computer Lab.

Living Room As shown in Fig. 5(a), we used a living room that would be typical of many homes. The living room presents itself as an ideal environment for replays as it is typical of where many people would place their Smart Speaker (VCD). The large space with carpeting would minimize any reverberation. Noise from air circulation systems may still occur spontaneously but at a much quieter level than in an office building. *Home Office Desk* As shown in Fig. 5(b), we used a home office desk as a playback environment. This would be a typical space where a person with several Smart Speakers (VCDs) in a home might place one. The hard surface of the desk and nearby walls will provide for some audio reverberation. *Copy Room* As shown in Fig. 5(c), we used a copy room located in a small alcove. This environment had some non-stationary noise caused by a copy machine and people walking by. *Mini Audio Booths* As shown in Fig. 5(d), we designed the mini audio booths to eliminate background sounds, i.e. sound of computer fans and air conditioners. Producing replay recordings in the audio booths allowed us to analyze the audio signals without ambient noise. The mini audio booths are effective in reducing ambient environmental noises. We used a sound pressure level (SPL) meter to determine an SPL of 40 dBA for a quiet office and 35.8 dBA inside the mini audio booth. Further testing shows that a large fan running 40 in. away from the microphone would produce an SPL of 56.5 dBA and 40.8 dBA in the office and audio booth, respectively. Moreover, running a vacuum cleaner produces an SPL of 70 dBA in the office and 42.5 dBA inside the audio booth. More specifically, we created two audio booths, one each for 1PR recordings and 2PR recordings. Shown in Fig. 5 (d) is the setup at 1PR, where a Bluetooth speaker is replaying an original audio sample. The Amazon Echo Dot is set to the "Drop-In" audio-conferencing mode so that it can replay the audio to another Echo Dot in the second audio booth. *Conference Room* As shown in Fig. 5(e), we used a conference room as it allowed for a relatively quiet space. This space still had noise from the air circulation system and footsteps from a nearby corridor. The narrowness of the room and the large surface of the table also allowed for the audio to reverberate. *Computer Lab* The computer lab playback environment used two computer labs adjacent to each other. One computer lab, shown in Fig. 5(f) is used for conducting the 1PR replay, where we placed a speaker, microphones and the Echo Dot in Drop-In mode. In the next computer lab, we arranged the other Echo Dot along with the necessary microphone to capture the 2PR playback. The computer labs contain the noise of air circulation systems only.

3.5. Equipment used for recording and playback

For recording the bona fide 0PR source files and the 1PR and 2PR replays, we used several combinations of microphones and microphone interface devices. More specifically, we used 35 distinct microphones for audio recordings, some of these are having the same make and model. For details, see the metadata file included with the dataset. Shown in Table 4 are the combinations of external microphones and their microphone interface devices. Devices with internal microphones are mentioned by the device name. Several professional-grade microphones that use an XLR connection are used for recording and playback. These microphones are connected to the PC using an audio interface box. The professional microphones connected by XLR cables are highlighted in green in Table 4. An external USB microphone is used for making some of the 0PR recordings. This type of microphone can be characterized as a medium quality microphone and is highlighted in yellow in Table 4. We also used various internal microphones of laptops and cell phones. Internal microphones can be characterized as lower quality microphones and are highlighted in red in Table 4.

Shown in Table 5 are the 14 1PR playback configurations used in the proposed VSDC. The composition of configurations consists of a speaker, amplifier, and a sound card. We used a variety of speakers ranging from low to high quality. Devices such as laptops that contain built-in speakers represent the low quality, whereas those using an external speaker either connected via Bluetooth or aux cable are considered devices of medium quality. Finally, the speakers considered to be high quality are those whose manufacturer's specifications report that they produce a sound frequency response near the full range of human hearing of about 80Hz - 20,000Hz.

Table 4
List of all configurations of recording microphone models and sound cards used.

0PR Recording Configuration
Audio-Technica ST95MKII | Zoom R16
Audio-Technica ST95MKII | Presonus Studio 24
Shure SM58 | Zoom R16
Shure SM58 | Presonus Studio 24
Behringer ECM8000 | Zoom R16
Electro-Voice 635A/B | Zoom R16
Blue Yeti | Mac Book Pro-2018
MacBook Pro-2018 (Internal Microphone)
Acer (Internal Microphone)
Samsung Galaxy S7 (Internal Microphone)
iPhone 5S (Internal Microphone)
iPhone 8 (Internal Microphone)
iPhone XR (Internal Microphone)
iPhone 7 (Internal Microphone)
(7) Android Phones
1PR and 2PR Recording Configuration
Audio-Technica ST95MKII | Zoom R16
Audio-Technica ST95MKII | Presonus Studio 24
Shure SM58 | Zoom R16
Shure SM58 | Presonus Studio 24
Behringer ECM8000 | Zoom R16
Acoustic Magic Array Microphone | Presonus Studio 24

xFormatted as [Microphone] | [soundcard] or [device]. Devices containing an internal microphone and soundcard are listed as the device name.

We used 8 different configurations of devices to transmit the 1PR audio having the corresponding 2PR recordings as shown in Table 6. The device configurations vary from Amazon Echos using the Drop-In audio conferencing feature to laptops and tablets connected using Google Meet.

Amazon Drop-In audio conferencing offers easy transmission of audio from one location to another using VCDs. We used several different combinations of Amazon Echo devices to test their audio quality. All of the Amazon devices, even the previous Generation 2 Echo Dot with smaller speakers, are able to replay commands to another VCD with acceptable results. All Amazon Echo devices contain audio-out jacks, which allow them to be connected to external speakers for improved quality. We connected external speakers for various sample sets to analyze the changes in the audio signal. We used a variety of external speakers ranging from small battery-powered external speakers to home theater speakers and studio monitor speakers.

For devices other than Amazon Echo, we used Google Meet as a means to transmit the 1PR audio to 2PR. Although Google Meet is only available on laptops and tablets, however, it can still be used to launch an audio replay attack. Laptops, tablets and phones can easily be left at other locations with the intention of replaying audio to VCDs at a later time. We used the laptops and Google Meet to ensure the use of high-quality microphones. The audio quality of the replays is limited to the quality of Amazon Echos microphone in the configurations where we used the Amazon Echo devices. It can be expected that competition amongst VCD manufacturers will continue to improve their audio capabilities, likely to the point that products will become available for

Table 5
List of all Playback configurations used in playing back audio in the 1PR recording environment.

1PR Playback Configuration
Polk R150 Speaker | Yamaha HTR-5840 | ZOOM R16
Polk R150 Speaker | Yamaha HTR-5840 | Asus GL504GM-DS74
Polk R150 Speaker | Yamaha HTR-5840 | USB Audio Card Ugreen 30521
Polk R150 Speaker | Fisher 143 | USB Audio Card Ugreen 30521
Bose 141 Speaker | Yamaha HTR-5840 | USB Audio Card Ugreen 30521
Presonus Eris E5 | ZOOM R16
Presonus Eris E5 | USB Audio Card Ugreen 30521
Bose Soundlink 415,859 | Asus GL504GM-DS74 (Wired)
Bose Soundlink 415,859 | Asus GL504GM-DS74 (Bluetooth)
SBT6050R | Asus GL504GM-DS74(Wired)
SBT6050R | Asus GL504GM-DS74(Bluetooth)
MacBook Pro-2018 (Internal Speaker)
Acer Nitro Spin 5 (Internal Speaker)
Acer Aspire E5-574G (Internal Speaker)

Table 6
VCD, 1PR to 2PR Replay Configuration.

Source	Target	Connection Method
Echo Dot 2	Echo Dot 2	Amazon Drop-In
Echo Dot 2	Echo Dot 3	Amazon Drop-In
Echo Dot 3	Echo Dot 2	Amazon Drop-In
Echo Dot 3	Echo Dot 3	Amazon Drop-In
Echo Dot 3	Echo Plus Gen 2	Amazon Drop-In
Echo Dot 3	Echo Input	Amazon Drop-In
LG G6	Asus Tablet	Google Meet
Laptop	Laptop	Google Meet

the audiophile market. Therefore, we conclude that it is worthwhile to study the audio characteristics of replay attacks on all types of audio equipment.

3.6. Data availability

The dataset is organized into different folders, where each folder has all the recordings (0PR, 1PR and 2PR) of a unique speaker. Each speaker folder contains three sub-folders, including the audio samples of 0PR, 1PR, and 2PR. The naming convention of the file specifies the sample number, point of replay, speaker, environment, microphone at 0PR, configuration number and phrase number as shown in Table 7. The proposed voice spoofing detection corpus is available at (VSDC, 0000) for research purposes. The elements of the file naming convention are described below.

- Sample set Number: Indicates the original 0PR file the sample is based from
- Point of Replay: Indicates at which point of replay this sample was created
- Speaker: Human speaker/volunteer who recorded this sample
- Environment: Recording Environment of the 0PR sample
- Microphone: The microphone that was used for the 0PR recording
- Configuration Number: The configuration setup that was used to make the 1PR and 2PR samples. The configuration is based on replay speaker, replay device, 1PR to 2PR transmission device and method
- Phrase: In each configuration, the volunteer spoke at least 9 phrases. This number indicates the phrase among all phrases spoke for that sample
- Example filename: 11-1PR-S2-E6-M7-C8-01

Shown in Table 8 is the number of samples collected at each point of replay. “Sample” refers to each 6-second audio file that can have any phrase listed in Table 3. The 0PR samples are the original bona fide phrases used. The 0PR files are replayed multiple times with unique microphone and speaker configurations to create the 1PR and 2PR samples.

The dataset contains 14,050 files indicated in Table 8. Many of the 0PR sets are used more than once to create the subsequent replay sets. We do not include these 0PR files more than once in the data set.

4. Experiments and results

We used standard spoofing countermeasure on multiple datasets to compare VSDC to three other datasets i.e. ASVspoof 2017 (Delgado et al., 2018), ASVspoof 2019 (Todisco et al., 2019), and ReMASC (Gong et al., 2019). Since constant Q cepstral coefficients (CQCC) features and Gaussian mixture model (GMM) classifier were used as a baseline of ASVspoof 2017 and 2019, we used the same for our experiments. For evaluation, ASVspoof 2017 used the metric of equal error rate (EER) while ASVspoof 2019 employed min-tDCF (tandem-decision cost function) along with the EER. Therefore, we used the EER metric for ASVspoof 2017 and EER and min-tDCF metrics (Kinnunen et al., 2018) for ASVspoof 2019 to perform comparative analysis with VSDC.

Table 7
An example of a replay configuration.

Example filename: 11-1PR-S2-E6-M7-C8-01	
Sample set Number	11
Point of Replay (PR)	1PR
Speaker	S2
Environment	E6
Microphone	M7
Configuration Number	C28
Phrase	01

Table 8
Number of samples for different points of replay.

Sample Type	Number of Samples
0PR	1687
1PR	6179
2PR	6184
Total	14050

4.1. Experiment-1: Training on proposed VSDC for multi-order replay attacks

To test the effectiveness of our dataset in terms of detecting both first- and second-order replay, we performed an experiment in two stages: first using bona fide and first-order replay samples, and then using bona fide and second-order replay samples.

In the first stage of this experiment, we evaluated the performance of the CQCC-GMM baseline countermeasure on our dataset using only bona fide and first-order replay samples. We used the 20-D CQCC variant, including the static, delta, and delta-delta coefficients. For this purpose, we used 60% samples for each of the bona fide and first-order replays to train the baseline countermeasure. In the second stage of this experiment, we evaluated the same baseline countermeasure using only the bona fide and second-order replay samples. The results obtained for both stages are shown in Table 9. Similarly, we repeated this experiment using the CQCC-GMM countermeasure using the 30-D CQCC variant including the static, delta, and delta-delta coefficients on VSDC. The results are reported in Table 9.

The results of this experiment indicate that first-order replays are more challenging to detect than second-order replay attacks, due to the fact that an additional playback device and microphone will increase distortion in second-order replays. Additionally, the baseline countermeasure using more CQCC features achieves a smaller EER as compared to the one using fewer CQCC features when evaluated on VSDC. More specifically, we observed a drop in EER of 3.25% on 0PR-1PR set and 1.96% on 0PR-2PR set. Therefore, we used the CQCC-GMM countermeasure with 30-D CQCC variant for all of the remaining experiments.

4.2. Experiment-2: Training on ASVspoof datasets

We designed a two-stage experiment to investigate the capability of the ASVspoof baseline CQCC-GMM countermeasure for replay detection in more diverse conditions. First, we trained the baseline CQCC-GMM countermeasure on the training samples of ASVspoof 2017 database version-2 (Delgado et al., 2018) and tested it on the development (dev) and evaluation (eval) sets of ASVspoof 2017 dataset, the 0PR-1PR and 0PR-2PR testing sets of the proposed VSDC (VSDC, 0000), and the testing set of ReMASC dataset (Gong et al., 2019). Similarly, in the second stage of this experiment, we trained the baseline CQCC-GMM countermeasure using training samples from the ASVspoof 2019 dataset. The results of this experiment are provided in Tables 10 and 11.

From the results (Table 10 and 11), we can clearly observe that the performance of the baseline countermeasure degrades on VSDC and ReMASC datasets. More specifically, we experience an increase in average EER of 26.62% on VSDC (both 0PR-1PR and 0PR-2PR sets) and 35.12% on ReMASC compared to the ASVspoof 2017 (dev + eval sets) dataset. Whereas, we observed an increase in EER of 7.74% on VSDC and 20.49% on ReMASC over ASVspoof 2019 (dev + eval sets) dataset. These results indicate

Table 9
Results of model trained on 0PR-1PR and 0PR-2PR Test sets.

Baseline Model		EER(%)	
Features	Classifier	0PR-1PR Test set	0PR-2PR Test set
CQCC 20-D variant	GMM	20.54	10.74
CQCC 30-D variant	GMM	17.29	8.78

Table 10
Performance evaluation on ASVspoof 2017, VSDC, and ReMASC.

	Dataset(Training)	ASVspoof 2017	ASVspoof 2017 & VSDC	ASVspoof 2017 & ReMASC
		EER (%)		
Dataset (Testing)	ASVspoof 2017 (Dev)	12.08	11.04	11.89
	ASVspoof 2017 (Eval)	29.95	25.24	27.04
	VSDC(0PR-1PR)	52.12	43.67	49.11
	VSDC(0PR-2PR)	43.16	32.17	39.98
	ReMASC (Test)	56.14	45.6	43.67

Table 11
Performance evaluation on ASVspoof 2019, VSDC, and ReMASC.

	Dataset(Training)	ASVspoof 2019		ASVspoof 2019 & VSDC		ASVspoof 2019 & ReMASC	
		EER (%)	min-tDCF	EER (%)	min-tDCF	EER (%)	min-tDCF
Dataset (Testing)	ASVspoof 2019 (Dev)	22.66	0.414	15.79	0.326	18.91	0.391
	ASVspoof 2019 (Eval)	23.16	0.424	17.39	0.347	21.89	0.402
	VSDC(0PR-1PR)	34.25	0.567	28.2	0.456	32.15	0.467
	VSDC(0PR-2PR)	27.05	0.504	22.5	0.407	25.98	0.487
	ReMASC (Test)	43.4	0.617	38.7	0.591	31.8	0.467

that the performance of the countermeasure has a dependency on the recording environment, and recording and playback settings.

4.3. Experiment-3: Training on combined set (ASVspoof and VSDC)

To test our hypothesis that training the anti-spoofing system with a more diverse dataset containing the attributes mentioned previously can enhance the performance of the baseline countermeasure, we designed our third experiment to train the CQCC-GMM baseline countermeasure on the combined corpus comprised of both the ASVspoof and our VSDC samples. We performed this experiment in two stages. First, we trained the baseline countermeasure on the combined corpus consisting of the training samples of ASVspoof 2017 and VSDC. Later, we tested the trained model on each of the three datasets and results are reported in Table 10. We observed a drop in EER of 4.71% on the eval set, 1.04% on the dev set, 10.9% on 0PR-1PR, 11% on 0PR-2PR, and 10.54% on the ReMASC test set than the EER obtained by the model trained only on the ASVspoof 2017 dataset.

In the second stage, we trained the CQCC-GMM baseline countermeasure on the combined training corpus of ASVspoof 2019 and VSDC and results are reported in Table 11. For ASVspoof 2019 dataset, we observed a drop in average EER and min-tDCF of 6.32% and 0.083 than those achieved on the model trained only on the ASVspoof 2019 dataset. Similarly, for VSDC and ReMASC datasets, we observed a decrease in average EER of 5.3% and 4.7%, and min-tDCF of 0.1 and 0.26 respectively.

From the results (Table 10 and 11) of this experiment, we can conclude that training the anti-spoofing model with additional audio samples collected in more diverse settings and scenarios can enhance the performance of the anti-spoofing model.

4.4. Experiment-4: Training on combined set (ASVspoof and ReMASC

To further investigate the effect of training the countermeasure on the combined corpus, we extend our third experiment to train the CQCC-GMM baseline countermeasure on the combined corpus consisting of both the ASVspoof and ReMASC samples. For this experiment, we partitioned the ReMASC dataset into “train” and “test” sets. Like the previous experiment, we also conducted this experiment in two stages. First, we trained the CQCC-GMM countermeasure on the combined corpus consisting of the training samples of ASVspoof 2017 and ReMASC and results are provided in Table 10. We observed a drop in EER of 0.19% and 2.95% on dev and eval sets, 12.47% on ReMASC, 3.01% on 0PR-1PR, and 3.18% on 0PR-2PR than the EER achieved on the countermeasure trained using the ASVspoof 2017 dataset only.

In the second stage, we trained the CQCC-GMM baseline countermeasure on the combined corpus of ASVspoof 2019 (train) and ReMASC (train) datasets and results are provided in Table 11. We observed a drop in average EER of 2.51% on ASVspoof 2019 dataset, 11.6% on ReMASC dataset, and 15.49% on VSDC than the EER obtained by the model trained only on the ASVspoof 2019 dataset. Hence, we conclude that training on this combined corpus improves the detection performance of the countermeasure.

From the results (Table 10 and 11) of this experiment, we can conclude that training the anti-spoofing model with additional audio samples collected in more realistic scenarios as reported in ReMASC (Gong et al., 2019) improves the performance of the anti-spoofing model.

4.5. Experiment-5: Analysis of spoofing detection performance on VSDC using different environments during the replays.

We designed this experiment to examine and compare the performance of the baseline countermeasure for spoofing detection on the samples replayed in different environments. For this purpose, we selected and categorized the samples from our dataset based on the environment used to generate the replays. For this experiment, we selected two rooms i.e. audio chamber and conference room. We selected the samples replayed in these two environments because they are different in terms of external noise. More specifically, the audio chamber is almost free of noise, whereas, conference room contains the noise of the air circulation system and footsteps. We performed this experiment in two stages. In the first stage, we evaluated the performance of the replay samples played in the audio chamber, whereas, in the second stage, we examined the replay samples in the conference room environment. For samples replayed in each of these environments, we divided our selected collection into two sets that are 0PR-1PR and 0PR-2PR. We used 60% of the samples from each collection for training and the rest for testing. The results are reported in Table 12.

From the results presented in Table 12, we observed that the samples replayed in the noise-free audio chamber obtained 5.3% and 3.99% higher (absolute) EERs on 0PR-1PR and 0PR-2PR test sets respectively, when compared to the samples replayed at the

Table 12
Analysis of spoofing detection performance on VSDC using different environments during replays.

Playback Environment	EER(%) 0PR-1PR Test-set	EER(%) 0PR-2PR Test-set
Audio Chamber	23.4	13.9
Conference Room	18.1	9.89

Table 13
Analysis of spoofing detection performance on VSDC using different playback devices during replays.

Playback Devices	EER(%)
Presonus Studio 24	20.47
Acer Aspire	16.15

conference room. This is attributed to the fact that the audio chamber does not add any environmental noise, so that only the artifacts caused by the playback device are present in the replay signal. Additionally, as per the patterns of results obtained in other experiments, we also observed low EER for 0PR-2PR than 0PR-1PR samples.

4.6. Experiment-6: Analysis of spoofing detection performance on VSDC using different playback devices.

We designed this experiment to examine and compare the performance of the baseline countermeasure for spoofing detection on the samples replayed with different devices. For this purpose, we selected the samples from our dataset based on the playback devices used for replays. For this experiment, we selected two playback devices i.e. Acer Aspire laptop with internal speaker (low-quality) and Presonus Studio 24 interface connected to a Presonus E5 external speaker (high-quality). We selected these two playback devices for this experiment due to their difference in playback quality. The results of this experiment are reported in Table 13.

From the results (Table 13), we observed that the samples replayed using the Presonus Studio 24 with high-quality Presonus E5 external speaker achieved a 4.32% higher EER than the samples replayed using the Acer Aspire a device with internal speakers. This is attributed to the fact that the low-quality built-in speaker of Acer Aspire device adds more artifacts in the replay signal that makes it more different from the bona fide recording and thus easier for the anti-spoofing model to classify between bona fide and replay samples.

4.7. Discussion.

VSDC aims to evaluate the performance of replay spoofing detection in multi-hop scenarios, when used in conjunction with other datasets such as ASVspoof (2017 and 2019) and ReMASC, it can also be used to evaluate the speaker verification and spoof detection in first-order replay scenarios. This section presents a performance analysis of different experiments (discussed above) on the proposed VSDC, ASVspoof (2017 and 2019), and ReMASC datasets using the CQCC-GMM countermeasure. The results of *Experiment-1* clearly indicate that second-order replays are easier to detect and achieve on average 9.15% smaller EER than the first-order replays.

The results achieved on the ASVspoof 2017 and ASVspoof 2019 datasets in *Experiments 2–4* show that the baseline countermeasure achieves lower EER on ASVspoof 2019 as compared to ASVspoof 2017 dataset. This pattern is visible in all these experiments (*Experiments 2–4*) regardles of whether the ASVspoof dataset is used alone for training or in combination with VSDC or ReMASC datasets. Perhaps this is because the simulation-based approach of ASVspoof 2019; introduces less distortion during the replay process compared to real replay recordings of ASVspoof 2017. The simulations are generated by changing one parameter at a time to examine different factors affecting the impact of replay attacks on ASV systems, along with the reliability of the anti-spoofing models. Additionally, our dataset configurations, environment and recording conditions are closer to the ASVspoof 2019 than ASVspoof 2017. This might be the possible reason of getting better results on VSDC for ASVspoof 2019 experiments (Table 11) than ASVspoof 2017 experiments (Table 10).

The results of *Experiment-5* indicate that the performance of countermeasures in terms of replay attack detection also depends on the playback environment, where significant background noise contributes to higher detection performance and vice versa. From the results of *Experiment-6*, we observed that the quality of speakers in playback devices also affects the detection performance of the countermeasure. More explicitly, devices with high quality speakers make it more difficult for the countermeasures to detect replay attacks.

5. Conclusion

VCDs can be exploited in IoT driven environments to generate various spoofing attacks including chains of replays. Existing audio spoofing datasets cannot be used to evaluate countermeasures against multi-hop voice spoofing in the IoT environment. Therefore, this paper presents a dataset consisting of bona fide, first-order, and second-order replay samples to evaluate anti-spoofing algorithms meant to detect multi-order replay attacks. The proposed dataset contributes to the existing spoofing datasets mainly through adding more diversity in playback scenarios (i.e. multi-hop replay attack), recording environments, use of array microphones along with non-array microphones (e.g. professional-grade microphones) at both 1PR and 2PR to capture audio characteristics useful to the development of future IoT applications.

Performance evaluation on CQCC-GMM countermeasure using the proposed VSDC shows lower EER compared to ASVspoof 2017 and ASVspoof 2019 datasets due to one or more of the following potential reasons: the use of microphone arrays along with non-array microphones (e.g. professional grade microphones) at both 1PR and 2PR, add the diversity in playback devices and acoustic environments. Evaluation also showed that the discrimination between bona fide and first-order replay samples is more challenging than discrimination between bona fide and second-order replay samples. Moreover, we conclude that the characteristics of playback devices must also be thoroughly investigated to identify the difference in features of the first-order replay and second-order replay spoofing samples. Additionally, results of training on combined sets (VSDC and ASVspoof) shows that training the anti-spoofing model with additional audio samples collected in more diverse settings and scenarios can improve the performance of the anti-spoofing model, therefore, VSDC could be used along with ASVspoof for the performance evaluation of future anti-spoofing methods.

Declaration of Competing Interest

We have no conflict of interest to declare.

Acknowledgement

This work was supported by a grant from the National Science Foundation (NSF) of USA via Awards No. (1815724) and (1816019).

References

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., Li, H., 2019. Protecting world leaders against deep fakes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 38–45.

Aircrack-Ng, 2019. https://www.aircrack-ng.org, accessed on August 8.

Alvi, S.A., Afzal, B., Shah, G.A., Atzori, L., Mahmood, W., 2015. Internet of multimedia things: vision and challenges. Ad Hoc Netw 33, 87–111.

Amazon has set its sights on healthcare tech with a stealth lab it calls '1492', 2019. https://www.businessinsider.com/report-amazon-pursues-healthcare-tech-1492-lab-2017-7, accessed on June 30.

Application, A. S., 2019. Available on https://www.audacityteam.org, accessed on August 15.

Chaudron, S., Geneiatakis, D., Kounelis, I., Di Gioia, R., 2019. Testing internet of toys designs to improve privacy and security. In: Mascheroni, G., Holloway, D. (Eds.), The Internet of Toys. Studies in Childhood and Youth. Palgrave Macmillan, Cham.

https://www.theguardian.com/technology/2017/nov/14/retailers-urged-to-withdraw-toys-that-allow-hackers-to-talk-to-children, 2020. accessed on June 14

CSTR VCTK Corpus, 2020. Available on https://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html, accessed on June 11.

ASVspoof 2017 dataset, 2019a. Available on https://datashare.is.ed.ac.uk/handle/10283/3055, accessed on June 29.

ASVspoof 2019 dataset, 2019b. Available on https://datashare.is.ed.ac.uk/handle/10283/3336, accessed on June 29.

Delgado, H., Todisco, M., Sahidullah, M., Evans, N., Kinnunen, T., Lee, K., Yamagishi, J., 2018. ASVspoof 2017 version 2.0: meta-data analysis and baseline enhancements. The Speaker and Language Recognition Workshop, Les Sables dOlonne, France. 296−303.

CVE-2019-3948, 2019. Detail https://nvd.nist.gov/vuln/detail/CVE-2019-3948, accessed on August 15.

https://www.bbc.com/news/av/technology-38966285/how-hackers-could-use-doll-to-open-your-front-door, on June 16, 2020. 2020.

Gong, Y., Yang, J., Huber, J., MacKnight, M., Poellabauer, C., 2019. ReMASC: realistic replay attack corpus for voice controlled systems. arXiv preprint: 1904.03365.

First the Home Now the Office., 2019. Amazon Launches Alexa for Business Service https://voicebot.ai/2017/12/01/first-home-now-office-amazon-launches-alexa-business-service, accessed on August 15.

Kinnunen, T., Lee, K. A., Delgado, H., Evans, N., Todisco, M., Sahidullah, M., Reynolds, D. A., 2018. t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification. arXiv preprint: 1804.09618.

Kinnunen, T., Sahidullah, M., Falcone, M., Costantini, L., Hautamki, R.G., Thomsen, D., Evans, N., 2017. Reddots replayed: a new replay spoofing attack corpus for text-dependent speaker verification Research. In: Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5395–5399.

Malik, K.M., Javed, A., Malik, H., Irtaza, S.A., 2020. A light-weight replay detection framework for voice controlled iot devices. in IEEE Journal of Selected Topics in Signal Processing. https://doi.org/10.1109/JSTSP.2020.2999828.

Malik, K.M., Malik, H., Baumann, R., 2019. Towards Vulnerability Analysis of Voice-driven Interfaces and Countermeasures for Replay Attacks. In: Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, pp. 523–528.

RedDots Project, 2020. Available on https://sites.google.com/site/thereddotsproject/reddots-challenge, accessed on June 11.

Sahidullah, M., Delgado, H., Todisco, M., Kinnunen, T., Evans, N., Yamagishi, J., Lee, K.A., 2019. Introduction to voice presentation attack detection and recent advances. In Handbook of Biometric Anti-Spoofing. Springer, Cham, pp. 321–361.

Todisco, M., Delgado, H., Evans, N., 2017. Constant q cepstral coefficients: a spoofing countermeasure for automatic speaker verification. Computer Speech & Language 45, 516–535.

Sugawara, T., Cyr, B., Rampazzi, S., Genkin, D., Fu, K., 2019. Light commands: Laser-based audio injection attacks on voice-controllable systems.

Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Lee, K. A., 2019. Asvspoof 2019: Future horizons in spoofed and fake audio detection. arXiv preprint: 1904.05441.

Voice Spoofing Detection Corpus (VSDC), http://www.secs.oakland.edu/~mahmood/datasets/audiospoof.html.