Practical Speech Re-use Prevention in Voice-driven Services

Yangyong Zhang
Texas A&M University
yangyong@tamu.edu

Sunpreet S. Arora
Visa Research
sunarora@visa.com

Maliheh Shirvanian
Visa Research
mshirvan@visa.com

Jianwei Huang
Texas A&M University
hjw@tamu.edu

Guofei Gu
Texas A&M University
guofei@cse.tamu.edu

ABSTRACT

Voice-driven services (VDS) are being used in a variety of applications ranging from smart home control to payments using digital assistants. The input to such services is often captured via an open voice channel, e.g., using a microphone, in an unsupervised setting. One of the key operational security requirements in such setting is the freshness of the input speech. We present AEOLUS, a security overlay that proactively embeds a dynamic acoustic nonce at the time of user interaction, and detects the presence of the embedded nonce in the recorded speech to ensure freshness. We demonstrate that acoustic nonce can (i) be reliably embedded and retrieved, and (ii) be non-disruptive (and even imperceptible) to a VDS user. Optimal parameters (acoustic nonce's operating frequency, amplitude, and bitrate) are determined for (i) and (ii) from a practical perspective. Experimental results show that AEOLUS yields 0.5% FRR at 0% FAR for speech re-use prevention upto a distance of 4 meters in three real-world environments with different background noise levels. We also conduct a user study with 120 participants, which shows that the acoustic nonce does not degrade overall user experience for 94.16% of speech samples, on average, in these environments. AEOLUS can therefore be used in practice to prevent speech re-use and ensure the freshness of speech input.

CCS CONCEPTS

• **Security and privacy → Embedded systems security**; **Domain-specific security and privacy architectures**.

KEYWORDS

voice-driven service, voice assistant security, nonce embedding, replay attacks

ACM Reference Format:
Yangyong Zhang, Sunpreet S. Arora, Maliheh Shirvanian, Jianwei Huang, and Guofei Gu. 2021. Practical Speech Re-use Prevention in Voice-driven Services. In *24th International Symposium on Research in Attacks, Intrusions and Defenses (RAID '21), October 6–8, 2021, San Sebastian, Spain.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3471621.3471855

RAID '21, October 6–8, 2021, San Sebastian, Spain

ACM ISBN 978-1-4503-9058-3/21/10…$15.00
https://doi.org/10.1145/3471621.3471855

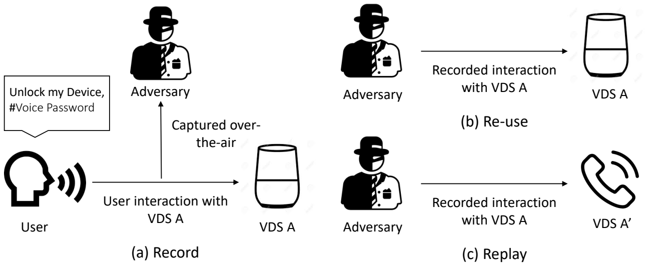


Figure 1: Difference between re-use and generic speech replays. In both cases, an adversary records a user's interaction with a voice driven service (VDS) "A", over-the-air as shown in (a). In case of re-use (b), the recorded interaction is used to attack the same type of VDS A, while in case of replay (c), the interaction can be used to attack the same or different type of VDS A'.

1 INTRODUCTION

Voice-driven services (VDS) are widely deployed in commercial products to enable personalized and convenient experiences for consumers. Examples include digital assistants from Amazon, Apple and Google for performing tasks like smart home control, hands-free online commerce, as well as automated voice response systems for customer support. While this increasing ubiquity can be primarily attributed to improved real world performance of deep learning driven speech and speaker recognition, security is an equally important consideration in operational settings. To secure VDS in practice, one of the common security mechanisms used is "voice password". For example, applications such as Samsung Bixby, Wechat [13, 19, 20] prompt users to speak a password and perform two factor authentication by checking (i) if the spoken passphrase is correct, and (ii) it was spoken by the authorized user.

Voice passwords, however, do not provide security against voice spoofing techniques such as speech replays [34, 43, 47] and synthesis [3]. While there are known practical limitations in conducting speech synthesis attacks (e.g., the attack proposed in [3] requires 24 hours of high quality training data), it is easier to record and replay user speech in practice. In this work, we aim to address a subset of malicious speech replays, called *speech re-use* attacks. We define speech re-use attacks to be specific to a certain type of VDS. As shown in Figure 1, let A and A' be the VDS types for the capture and attack steps, respectively. Speech replays typically consider a variety of VDS types for capture and attack steps (A = A' and A ≠ A'), while re-use attacks are a subset of such replays where the VDS types in consideration for the capture and attack steps is similar (A = A'). An example of a re-use scenario is replaying a recorded

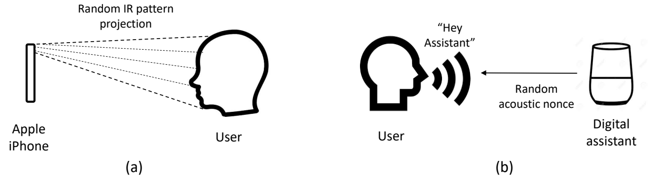


Figure 2: (a) Face ID's device-specific random IR pattern projection at the time of user authentication for preventing re-use [5]. Analogous to (a), the proposed security overlay (b) uses a random acoustic nonce at the time of user interaction to prevent speech re-use in voice-driven systems.

passphrase to fool the authentication process of an Amazon Alexa service, and then conduct online shopping or controlling a smart home device. A generic speech replay assumes this passphrase can be used for other VDS which is not a practical assumption because such passphrases are supposed to be VDS-specific or device-specific. For example, reciting the phrase "Hey Siri" will not trigger an Amazon Alexa service.

Existing defense mechanisms can be categorized into two classes. The first class of methods use machine learning to determine whether a speech sample is produced by a human or replayed using a playback device [26, 57]. The second category of techniques use external devices (e.g., a vibration sensor) to check if the speech is produced by a human in real-time [32]. For machine learning techniques, real world performance usually depends on training data and its relevance to the test environment. A state-of-the-art method Void [22], as an example, has significantly different equal error rates (EER), 0.3% and 11.6%, on two speech datasets collected in two different environments. Another practical limitation is unreliability of playback device signatures used by machine learning techniques [57]. In addition, for methods that use extra hardware, there are usability and cost implications for practical deployment.

We present AEOLUS[1], a security overlay to prevent re-use attacks on protected VDS in a proactive and device-agnostic manner. The proposed security overlay proactively embeds a dynamic acoustic nonce in the voice channel via the loudspeaker at the time of user interaction, and detects the presence of the embedded nonce in the speech recorded by the speaker to ensure speech freshness. For example, if the security overlay is integrated with each Alexa's VDS, it can prevent speech captured from user interaction with one Alexa's VDS from being re-used on another Alexa's VDS. This is similar to the device specific random pattern used in Face ID to counter digital and physical spoofs [5] (see Figure 2). AEOLUS is designed as a software solution, and can be integrated by a vendor with any closed-loop VDS that involves real-time user interaction. It is useful for providing additional security in critical applications such as user authentication and payments. AEOLUS does not require any extra hardware and works with in-built speaker and microphone in consumer products. AEOLUS is secure by design against attacks that remove the embedded nonce for speech re-use. It uses Frequency-hopping spread spectrum (FHSS) [29] [30] technique for dynamic acoustic nonce embedding. Similar techniques have been previously used to watermark audio recordings for copyright purposes [53] [23]. However, their application in offline audio watermarking is fundamentally different from VDS operational setting where the length and content of the input speech is unknown apriori. AEOLUS is designed for real-time over-the-air use without prior knowledge of the input speech.

AEOLUS addresses two key practical challenges. The first challenge is to reliably embed an acoustic nonce over-the-air and retrieve it successfully from the recorded speech. Over-the-air acoustic nonce propagation is affected by factors such as background noise and distance between microphone and loudspeaker [27]. The latter, for instance, results in distorted and attenuated speech signal which increases the likelihood of errors while extracting the embedded acoustic nonce. The second challenge is to embed the acoustic nonce such that it is imperceptible to a VDS user. Achieving imperceptibility is non-trivial in light of the first challenge. This is because an important consideration for reliability is embedding acoustic nonce of certain strength which, in turn, makes the acoustic nonce perceptible. Therefore, we model acoustic nonce generation as an optimization problem and compute the set of optimal parameters, e.g., nonce's operating frequency, amplitude, bitrate in different environments using differential evolution.

We evaluate AEOLUS in three different environments and show that it works reliably upto a range of 4 m. Additionally, we conduct a user study involving 120 subjects (approved by the institutional review board, or IRB) to evaluate the imperceptibility of the embedded nonce. The results of the study show that majority of users find the embedded acoustic nonce to be either imperceptible or non-disruptive.

In summary, the contributions of this work are as follows:

- Design of a VDS security overlay called AEOLUS to ensure freshness of input speech without any extra hardware dependency.
- Modeling acoustic nonce propagation as an optimization problem to address i) reliability, i.e. ensuring successful embedding and retrieval of dynamic acoustic nonce without impacting VDS functionality, and (ii) imperceptibility, i.e. to have minimal impact on VDS users' experience.
- Comprehensive real-world evaluation to show that AEOLUS can work effectively (0.5% false reject rate (FRR) and 0% false accept rate (FAR) in detecting speech re-use) up to a range of 4 meters in three different environments.
- User study with 120 subjects to demonstrate that the embedded acoustic nonce does not degrade user experience.

2 BACKGROUND AND RELATED WORK

In this section, we first outline typical VDS workflow for the collection and processing of a user's voice. Next, we discuss key motivating examples to underscore the need for designing AEOLUS. Following this, we review related work for speech re-use prevention and related audio watermarking methods.

2.1 Voice-driven Services (VDS)

VDS are becoming increasingly popular for conducting a variety of tasks ranging from internet browsing to online commerce. In fact, it is predicted that by the end of 2020, 50% of all searches performed online will be driven by voice [7]. VDS are widely used in both

[1] Named after Aelous, the "Keeper of the Winds" in Greek mythology.

smartphone and Internet of Things (IoT) platforms. Major smartphone vendors (e.g., Apple and Samsung) provide native support for VDS, e.g. Amazon's Alexa or Apple's Siri, in their operating systems, and make them available as independent applications. These can be used to instruct third-party applications to perform specific tasks, e.g., to turn lights on/off.

A VDS is operated by a user using voice commands. The voice command is processed by the service and a response is returned either verbally to the user (e.g., announcing the time) or by performing a specific action (e.g., turning lights on/off). Devices that use VDS are triggered by a user using a pre-specified phrase (e.g., Alexa, Hey Siri), and operate in a turn-taking mode. The device actively listens to the user once triggered and the user listens to the device when it responds. AEOLUS embeds the acoustic nonce once the device is triggered and is actively listening to the user.

A typical VDS has four different sub-components, (i) an input/output voice channel to listen to a user's command and send verbal responses, (ii) a text/speech processing engine to convert speech-to-text and text-to-speech, (iii) a speaker verification/identification module to only permit authorized users to execute privileged commands, and (iv) a text/intent processing engine to execute commands. AEOLUS aims to prevent speech re-use of an authentic VDS user.

2.2 Example Use Cases

Voice-driven Payments. Voice-driven services are being increasingly used for payments, especially in retail environments [14]. Amazon and Exxon recently announced voice-driven payments at over 11,500 gas stations in the US [2]. A recent user study [11] indicates that while most users are comfortable conducting low-value purchases such as ordering a meal and shopping for groceries using voice-driven payment services, the majority of them do not have sufficient confidence in conducting high value purchases due to security concerns.

Most voice-driven payment services authenticate users through speaker verification. However, they often lack adequate protection mechanisms against speech re-use attacks where an attacker records a user interaction and re-uses it to attack the service either on-site or remotely. This is analogous to card skimming [48] where an attacker re-uses a victim's credit card information. AEOLUS can be deployed to prevent speech re-use attacks in voice-driven payments.

Workplace Automation. A variety of digital assistants are being used for unsupervised or semi-supervised interactions in workplace environments [1]. In addition to accurate speaker recognition (identification and/or verification), it is important to have adequate protection mechanisms against speech re-use to protect sensitive resources in such environments [15]. AEOLUS can be incorporated in workplace digital assistants for this purpose.

2.3 Related Work

Speech Replay Detection Existing research related to speech replay detection can be classified into two categories, software-based and hardware-based methods. Software-based methods use machine learning to determine whether the input speech is produced by a human or replayed using a recording device [26] [22] [38]. Chen et al. [26] present a method that obtains 0% EER for distance (between user and VDS input device) of only a few centimeters which limits its applicability in practice. Ahmed et al. [22]'s method, on the other hand, is reported to work well upto a distance of 2.6 meters. However, the EER varies significantly in different environments with it being as low as 11.6%. One of the best performing methods [38] reports EER of 6.7% on ASVSpoof 17 database. A caveat, however, is that these methods aim to address general replay attacks. In contrast, AEOLUS aims to prevent speech re-use and obtains 0% EER upto a distance of 4 meters in three different environments.

The second category of approaches use additional hardware (e.g., a vibration sensor on user's neck) to check that speech is produced in real-time by a user [32] [47]. Such approaches, in general, outperform machine learning-based methods. For example, Feng et al. [32] report 97% successful detection rate, on average, with no strict assumption on distance. However, there are two major practical limitations using the aforementioned approaches for detecting speech re-use, (i) requirement of extra hardware which has cost implications, and (ii) inconvenience to users. AEOLUS only uses the built-in loudspeaker and microphone in smart devices yet achieves high speech re-use detection rate.

Audio Watermarking. Audio watermarking is widely used for multimedia copyright protection [49, 53]. Both time domain [23] and frequency domain [28, 44] watermarking techniques have been proposed in the literature. Typically, audio watermark embedding and detection is self-synchronized [59]. Advanced watermarking methods use spread-spectrum modulation [53] to prevent watermark removal and provide imperceptibility [28, 39]. While useful, these techniques cannot be directly used in VDS operational setting. They are designed for fixed length offline audio files, e.g., music recordings in CDs [53], and do not consider the impact of environmental factors, e.g., bitrate variability and background noise, in over-the-air transmission. Such environmental factors result in a dynamic and lossy acoustic environment in turn making watermark embedding and retrieval significantly challenging. Further, they are not designed for real-time speech and speaker recognition systems where input speech is unknown apriori and can be of variable length. AEOLUS is designed to take these pragmatic considerations into account. It is able to successfully embed and retrieve acoustic nonce in the voice channel without degrading user experience in practice.

3 PROPOSED SECURITY OVERLAY

In this section, we present the AEOLUS design by first introducing its threat model including both adversary capability and hardware assumptions. Next, as shown in Figure 3, we illustrate the core design of AEOLUS which consists of VDS-compatible acoustic nonce generation, embedding, and detection.

3.1 Threat Model

At the time of user interaction, the method used to trigger a VDS, e.g., words or phrases like "Alexa" and "Hey, Siri" are not protected but can initiate the proposed overlay to prevent re-use of voice

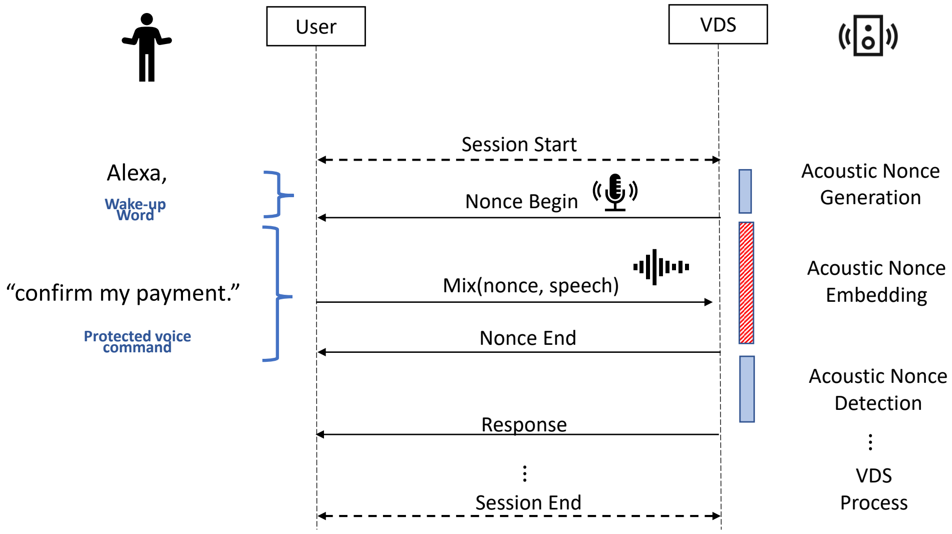


Figure 3: Use of the proposed security overlay to prevent speech re-use in voice-driven services. At the time of user interaction, an acoustic nonce is generated and embedded in over-the-air channel, and its presence is detected in the recorded speech to ensure speech freshness.

commands and confidential passphrases. For example, to prevent re-use of a user's voice password to access recent account transactions when the user is interacting with a voice banking VDS [18].

The following assumptions are made while designing AEOLUS to address the speech re-use attack scenario.

- **Adversary Capability.** We aim to address speech re-use in the context of VDS protected with AEOLUS. An adversary uses a commodity recording device to record prior user interactions with a AEOLUS-enabled VDS, and re-use the recorded interaction to attack a AEOLUS-enabled VDS. Hence, attack scenarios where an adversary can record or access clean user speech samples (i.e., recordings without embedded acoustic nonce) from a different context to launch replay or speech synthesis attacks are not considered. The key underlying assumptions are that (i) users typically interact with VDS using specific commands (e.g., a user-defined passphrase [20]) and these commands can be protected with AEOLUS, and (ii) it is difficult to record clean samples of these commands from interactions with different VDS without AEOLUS protection or in human conversations. Also, it is assumed that AEOLUS can be used in conjunction with other defense mechanisms for replay or speech synthesis attacks.
- **Hardware Assumptions.** We study the problem of speech re-use prevention wherein speech is presented over-the-air. It is assumed that the hardware device (loudspeaker and microphone) used by a user to interact with VDS is trusted, secure and functioning, and an adversary cannot disable the speaker or the microphone. Furthermore, we assume that the channel between hardware and other VDS modules is secure/protected. Similar hardware assumptions are made by other software-based security mechanisms such as CAPTCHA. The proposed framework does not address hardware integrity which can be addressed using other security mechanisms, e.g., hardware attestations.
- **Protection Scope.** An adversary who has access to the proximity of an AEOLUS-enabled device may inject audible or inaudible voice commands [21, 61] during users' interaction with the devices. This type of attack may happen before or during the period when users speak voice commands. AEOLUS does not provide protection again such malicious voice command injection because the goal of AEOLUS is to prevent speech re-use rather than providing the integrity of the current speech.

3.2 Core Components

To prevent speech re-use in VDS, AEOLUS uses three core components, nonce generation, nonce embedding and nonce detection.

Nonce Generation. The nonce generation module is invoked upon initiation of a new user interaction session (see Figure 4 ①). Akin to nonce-based authentication [24, 55], the module generates a digital nonce (a random number of a length L) that is valid for the current session. Nonce length L should be sufficiently large to prevent use of duplicate nonces in different sessions. Given limited bandwidth of over-the-air channel and estimated user base for popular VDS (e.g., Alexa skill has 300,000 daily users [16]), the minimum permitted value of L is set to 32 bits. 32-bit long nonce is sufficient to protect short speech samples (e.g., a voice passphrase) in a particular session. Nonce repetition is permitted to protect longer speech samples during the session. The nonce δ for session S is stored in a set S_{set}. S_{set} is subsequently used by the nonce detection module to determine if the embedded nonce is current.

Nonce Embedding. Post nonce generation, an encoding algorithm, Binary Frequency Shift Keying (BFSK) [45, 54], is used to encode the generated digital nonce as an acoustic nonce. Each bit is independently encoded as a cosine wave ω_i. Bit "0" is encoded as a 1kHz cosine wave, and bit "1" as a 5 kHz cosine wave. All cosine waves in the set ω are concatenated to yield the acoustic nonce δ. The acoustic nonce δ is embedded in over-the-air channel using

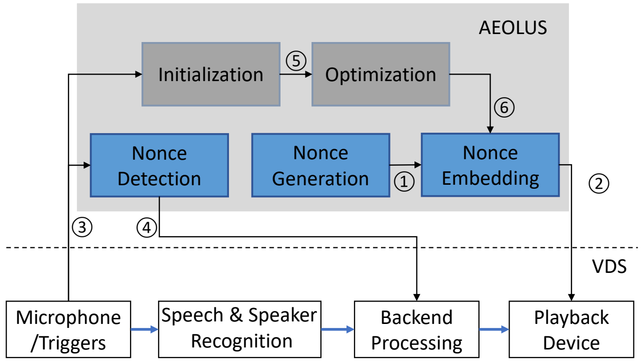


Figure 4: AEOLUS-enabled VDS components with the associated data flows. Shown in blue and grey are the components used to prevent speech re-use.

VDS playback device (Figure 4 ②). An important parameter for nonce embedding is the time duration of the embedded acoustic nonce relative to the time duration of user interaction. Using a real-world dataset with 27,000 samples [35], we estimate that the average user interaction is 3 sec. As an example, let the minimum duration of embedded nonce be half of the average user interaction. In this case, the acoustic nonce embedding would require a bitrate of 14 bits/sec, and the maximum duration of each component cosine wave w_i can be 42 ms. Note that this estimate does not consider the idle time in user interaction. Additionally, to prevent an adversary from obtaining clean speech sample by removing the embedded nonce, the nonce embedding module leverages Frequency-hopping Spread Spectrum (FHSS) [29, 30]. FHSS is a well-known technique that provides high robustness against standard embedded signal removal attacks. For high robustness, the module uses FHSS with a set of different frequencies and periodically selects the operating frequency at random.

Nonce Detection. Once user interaction ends, the recorded audio is processed using the nonce detection module (Figure 4 ③). The module decodes the acoustic nonce using BFSK, and checks that (i) the nonce is completely decoded, and (ii) the decoded nonce is current. The recorded speech is subsequently processed using standard VDS modules (Figure 4 ④). If nonce detection fails or if the decoded nonce has a high Bit Error Rate (BER), a recourse action (e.g. a retry or reject) can be taken. Furthermore, if speech is re-used by an adversary and both the current nonce and the obsolete nonce from re-used speech are detected, information about re-used speech such as when and where it was recorded can be determined from S_{set}. Another scenario can be signal interference between the current and obsolete nonce resulting in inappropriate decoding result. In such case, pertinent recourse action can be taken.

4 PRACTICAL REALIZATION

In this section, we first illustrate two key challenges while implementing AEOLUS in practice: reliability and imperceptibility. Then, we discuss the root causes of these challenges. Lastly, we present an optimization scheme to tackle the challenges.

4.1 Challenges

Reliability. The first challenge is to reliably embed an acoustic nonce in over-the-air channel, as well as to accurately detect nonce from the recorded speech. Because over-the-air acoustic nonce transmission is impacted by factors such as background noise and distance between microphone and playback device, it is difficult to achieve this in operational settings. It is also important to determine if any previously embedded acoustic nonce is present in the recorded speech to ascertain speech re-use. The metric used to measure reliability is the bit error rate (BER) between the embedded and detected nonce. Recall that BFSK is used for acoustic nonce embedding and detection. Under the assumption that only additive white Gaussian noise (AWGN) is present in over-the-air channel, BER can be computed using the complimentary error function $erfc()$ [40] as follows:

$$BER = \frac{1}{2} erfc(\sqrt{E_b/N_0}) \tag{1}$$

Note that BER computation in Eqn. 1 involves the normalized per bit signal-to-noise ratio E_b/N_0 [36] which depends on the frequency f and amplitude α of the carrier wave. However, other types of noises besides AWGN are typically present in over-the-air channel. Hence, we conduct Room Impulse Response (RIR) simulation experiments to study the impact of carrier wave amplitude and frequency on BER.

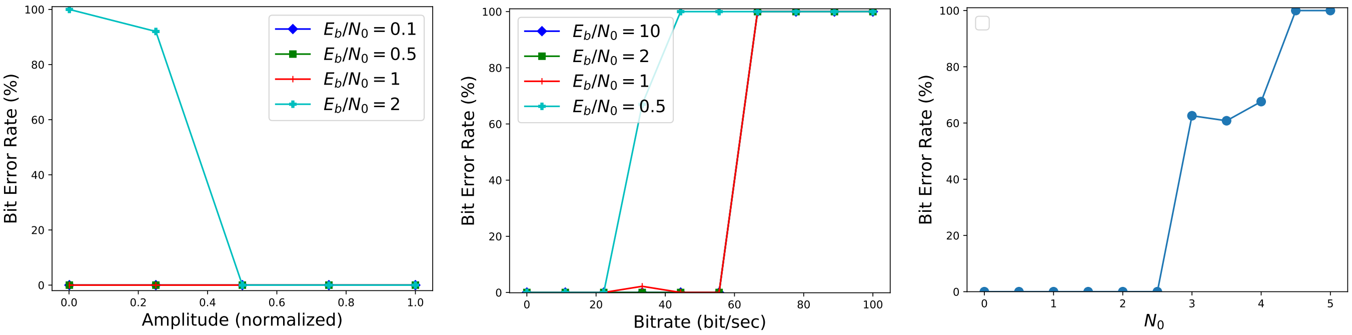
Imperceptibility. The second challenge is to ensure that the embedded acoustic nonce does not degrade VDS user experience. For this, the acoustic nonce should be as imperceptible as possible to a VDS user. The nonce embedding and generation modules presented earlier do not adequately address this challenge because (i) they do not optimize any objective metric for imperceptibility, and (ii) they do not account for dynamic and noisy environments where it is difficult to achieve both reliability and imperceptibility simultaneously. For measuring imperceptibility, Sound Pressure Level (SPL) is computed using the following equation:

$$SPL = 2\pi^2 f^2 \alpha^2 \rho c/v \tag{2}$$

Here, f represents the frequency and α is the amplitude of the carrier wave. ρ represents the density of the transmission channel (e.g., over-the-air channel), and c denotes the speed of acoustic transmission. Given that the average SPL for human speech is 60dBm [12], f and α should ideally ensure that SPL is below this threshold for imperceptibility. Like Equation 1, while this equation provides a theoretical basis to understand SPL, we study how the aforementioned parameters impact SPL using RIR simulation.

4.2 Key Parameters

There are two key parameters that impact reliability and imperceptibility: (i) acoustic nonce generation and embedding parameters that include frequency and amplitude of the carrier wave used for acoustic nonce generation and bitrate used for acoustic nonce encoding, and (ii) environmental parameters that include the distance between the microphone and user, the room size, and background



(a) Frequency range = 4-5kHz, hopping = 2, bitrate = 30 bits/sec

(b) Frequency range = 4-5kHz, hopping = 2, amplitude = 1

(c) Frequency range = 4-5kHz, hopping = 2, bitrate = 30 bits/sec, amplitude = 1

Figure 5: Variation of bit error rate (%) with: (a) amplitude of carrier wave, (b) bitrate used for encoding acoustic nonce, (c) background noise levels when embedding the acoustic nonce. E_b/N_0 is calculated in decibels. The distance between user (simulated using a loudspeaker) and the microphone is set to 1 m.

noise among others. To understand the impact of the these parameters on reliability and imperceptibility, we implement AEOLUS as a prototype and setup an RIR environment called MCRoomSim in MATLAB [56]. RIR is a commonly used tool for estimating the performance of signal processing applications in over-the-air channel [21, 27].

4.2.1 Nonce Generation and Embedding Parameters.

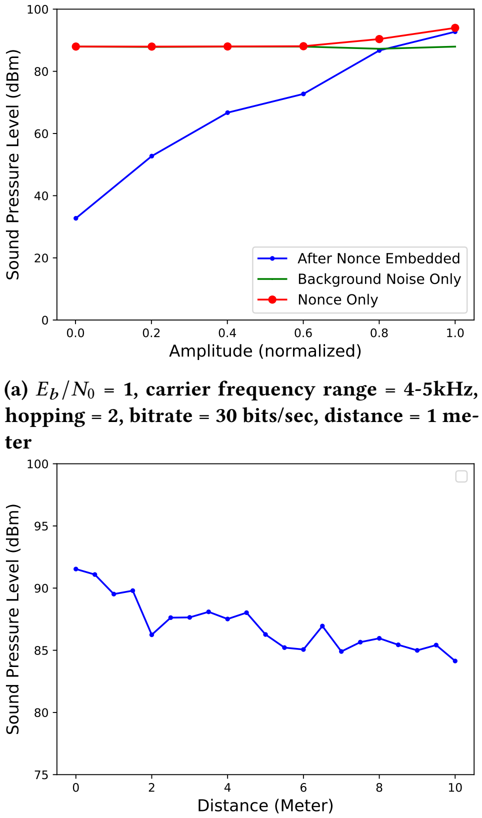
Amplitude. According to RIR simulation, a lower amplitude yields lower E_b/N_0 ratio and consequently, higher BER (see Figure 5a). This is consistent with Equation 1. Furthermore, as shown in Figure 6, SPL is proportional to the amplitude of the carrier wave.

Frequency. BER and SPL are evaluated in the frequency range at which typical speaker recognition systems operate (5 Hz - 8 kHz; sampling rate = 500 Hz). The other parameters are fixed as follows: $E_b/N_0 = 1$, frequency hopping = 2, amplitude (normalized) = 1, bitrate = 30 bits/sec, distance = 1 meter. It is observed that the carrier wave frequency neither impacts BER nor SPL. The later is in contrast to existing research [4, 52] which shows that the frequency of an acoustic wave affects the loudness as perceived by humans in a non-linear manner. Hence, the frequency to SPL mapping from IoSR toolkit [6] is used as reference.

Bitrate. Figure 5b shows the impact of acoustic nonce bitrate on BER. BER increases significantly as bitrate increases beyond 20 bits/sec. One of the causes is asynchronous signal overlapping in the time domain when two bits are transmitted too close to each other.

4.2.2 Environmental Parameters.

Frequency Shifting. Over-the-air acoustic nonce transmission induces differences in the received frequencies and the transmitted frequencies called *frequency shifting*. Frequency shifting can also occur because of hardware limitations, e.g., a microphone with a 1.8 kHz frequency limit being used to capture a 2 kHz wave. Since the received frequency is used to decode the acoustic nonce, frequency shifting needs to be accounted for by AEOLUS. We investigate frequency shifting in the simulation setting (operating frequency range: 500 Hz-8 kHz, cardioid loudspeaker type, dipole microphone type) and determine the frequency shifting range to be +/- 15 Hz.



(a) $E_b/N_0 = 1$, carrier frequency range = 4-5kHz, hopping = 2, bitrate = 30 bits/sec, distance = 1 meter

(b) $E_b/N_0 = 1$, carrier frequency range = 4-5kHz, hopping = 2, bitrate = 30 bits/sec, Amplitude (normalized) = 1

Figure 6: Variation of sound pressure level (dBm) with (a) amplitude of carrier wave, and (b) distance between user (simulated using a loudspeaker) and the microphone.

Distance. Distance between a VDS user and VDS input impacts imperceptibility of the acoustic nonce. Figure 6b indicates that this distance is inversely proportional to SPL.

Background Noise. Figure 5c and Equation 1 suggest that the background noise is proportional to BER. In an environment with high background noise, the nonce generation and embedding parameters need to be configured appropriately (e.g., decrease bitrate, increase amplitude) to limit the BER.

4.3 Mathematical Formulation

To address reliability and imperceptibility simultaneously, we model over-the-air acoustic nonce transmission as an optimization problem. Let a user's recited speech be denoted by x, and the acoustic nonce embedded by AEOLUS be δ. Also, let the function that accounts for variations induced by over-the-air channel, e.g., due to acoustic signal air absorption and distance, be $h(\cdot)$. The speech recorded by the microphone is thus represented as $h(x)$. When δ is embedded at the time of user interaction with the VDS, x mixes with δ. As a result, the microphone records $h(\mathbf{x}*\delta)$, where $*$ denotes signal convolution between x and δ.

Assume that AEOLUS detects nonce δ' in $h(x * \delta)$. Let us define an objective function $f(\cdot)$ denoting BER that takes the parameter set $\mathcal{P} = \{< f, \alpha, \tau >, ...\}$ as input, and a function $g(\cdot)$ denoting SPL. The goal is to find the optimal parameter set $\mathcal{P}$ that minimizes f (for reliability) subject to the constraint that the SPL difference between speech with and without embedded nonce is less than or equal to a threshold θ (for imperceptibility):

$$\arg\min_{\mathcal{P}} f(\mathcal{P}) \text{ subject to } g(h(\mathbf{x} * \delta)) - g(h(\mathbf{x})) \leq \theta \tag{3}$$

4.4 Computing Optimal Parameters

Initialization. The environmental parameters are first initialized (Figure 4 ⑤) using the VDS input. The background noise energy is calculated by inspecting the ambient sound. Frequency shifting is calculated by transmitting and decoding a nonce that covers the working frequency range (500 - 8 kHz, sampled every 500 Hz). It is assumed that the distance between the user and VDS input device is set by the VDS device vendor.

Optimization. Computing the gradients of the optimization parameters directly using gradient-based methods is inefficient [25]. This is because the VDS's acoustic environment is dynamic (e.g., due to occlusions and varying background noise levels). In addition, the formulated optimization is a multi-dimensional problem with considerations of both reliability and imperceptibility. Hence, we use differential evolution (DE) [31, 51] which is a gradient-free method. DE generates diverse candidate solutions in each iteration which avoids local minima and aids in convergence to a global minima. It does not require apriori knowledge of the underlying acoustic encoding method or acoustic environment. Algorithm 1 summarizes the use of DE for obtaining optimal parameters. The initial set of candidate parameters $\mathcal{P}$ includes feasible frequency f, amplitude α, and bitrate τ, estimated using RIR simulation. Once AEOLUS is deployed, the initialization module first estimates the environmental parameters, e.g., N_0 and d. Subsequently, the optimization module uses the pre-calculated RIR parameters for efficient convergence to optimal parameters (see Section 5.1).

Algorithm 1: Computing Optimal Parameters

```
initialization:
1. Create candidate population: $\mathcal{P} = \{P < f, \alpha, \tau >, ...\}$
2. Initialize: $\mathcal{P}_{output} = \{\}$, $i = 0$, $MAX_i = 100$,
   $MAX_{output} = 20$, $\beta = 0.5$
2. Acquire environmental parameters: $d$, $N_0$
3. Set up Goals: $BER < 10^{-9}$, $SPL' - SPL \leq \theta$
# $SPL$ : the average loudness value measured before the
nonce embedding; $SPL'$ : the average loudness value
measured during the nonce embedding
optimization:
while $i <= MAX_i$, $\mathcal{P}_{output}.Size <= MAX_{output}$ do
    foreach individual $P_i (i = 1, ...) \in \mathcal{P}$ do
        if $P_i$ satisfies the Goals then
            $\mathcal{P}_{output} \leftarrow P_i$
        end
        Create candidate $C$ from parent $P_i$, $d$, $N_0$.
        Evaluate the candidate $C$ and acquire $SPL_C$ and
         $BER_C$.
        if $SPL_C < SPL_{Parent}$ and $BER_C < BER_{Parent}$
         then
            $P_{feasible} \leftarrow$ Candidate $C$
            Candidate $C$ replaces the parent
        else
            the candidate is discarded
        end
    end
    Randomly select next $P$ in $\mathcal{P}$.
end
output: $\mathcal{P}_{output}$

Candidate Creation:
input :$P_{parent}$, $d$, $N_0$, $\beta$
Randomly select individuals $P_1$ , $P_2$ from $\mathcal{P}$
Calculate candidate $C$ as $C = P_{parent} + \beta(P_1 - P_2)$,
where $\beta$ is a scaling parameter decided by $d$ and $N_0$.
output:Candidate $C$
```

BER and SPL Calculation. The BER is calculated by comparing the generated nonce that is embedded over-the-air and the decoded nonce from the captured speech. With synchronization techniques [41] used in the nonce embedding and detection, AEOLUS is able to calculate how many error bits are received. For example, if bit sequence "0010" is transmitted and "0110" is received, BER is 25%. SPL is calculated using Equation 2. AEOLUS first calculates the frequency and amplitude of the captured speech. These parameters along with other constant values (e.g., density of transmission medium) are fed into Equation 2 to obtain SPL. In Algorithm 1, BER threshold of 10^{-9} is used because it is a commonly used threshold for a good BER in communication systems. Similarly, threshold value of 5% is used for determining small SPL difference.

Candidate Comparison. A candidate's parameter set is compared with its parent based on their performance, i.e., the resulting BER and SPL. If both the BER_C and SPL_C of a candidate is better than the parent's BER_{parent} and SPL_{parent}, the candidate is added to the feasible set and the parent is discarded. If not both of the BER_C or SPL_C of a candidate outperform BER_{parent} and SPL_{parent}, the candidate is discarded.

5 EXPERIMENTAL EVALUATION

AEOLUS is implemented as a prototype in MATLAB and evaluated in three different environments (see Figure 7): (i) a conference room (approx. 5.5 m by 4 m) selected as a reference for a semi-public indoor environment, (ii) a campus dining area (approx. 25m by 50m) chosen to mimic a noisy indoor public area, and (iii) inside a car parked at a gas station to simulate the use case of voice-driven payments at gas stations [2]. For (i) and (ii), a FIFINE K669B microphone is used to record speech, and a Logitech S120 Speaker System is used as playback device to simulate users. The evaluation is performed at different distances (0.5-4 m) between the microphone and playback device. For (iii), the microphone is identical but the car's in-built loudspeakers are used instead of the Logitech loudspeakers. Also, the evaluation is performed at two different distances 0.5 and 0.7 m.

AEOLUS is evaluated both on speech data captured from live participants in ReMASC dataset [35] and synthesized speech from Amazon Polly. A total of ten different speaker profiles are used (four of which are synthesized). For each profile, three different speech samples are used. Microsoft Azure's speech and speaker recognition services [9] are used for speech-to-text and speaker recognition. Text-independent speaker verification service [8] is used for speaker verification experiments. Each speaker's profile is first enrolled using a speech sample longer than 20 seconds. Verification is performed using the speaker's speech samples with and without acoustic nonce, and yields one of the following outputs: acceptance with "high", "normal", or "low" confidence, or a rejection.

5.1 Performance

Optimal Parameters. First, we estimate optimal parameters for achieving reliability and imperceptibility simultaneously in a given environment using Algorithm 1. The frequency step size is set to 200 Hz and θ is set as 0.1 to ensure that the generated nonce is below the average human speech SPL of 60 db. It is observed that lower amplitudes and frequencies and higher bitrates are comparatively optimal for acoustic nonce embedding in low noise environments, e.g., conference room (see Table 1). Relatively higher amplitudes and frequencies, and lower bitrates are more optimal in environments with higher noise, such as the dining hall. This is because the background noise in such environments typically vests as a low amplitude low frequency noise.

Computational Overhead. Algorithm 1 iteratively optimizes the parameter set $\mathcal{P}$ for the given environment. For conference room, a single iteration is sufficient. However, for dining hall and gas station environments, a few iterations (< 10) are required and the computation takes a few seconds. This does not cause a delay at the time of user interaction because it overlaps with the time a user typically takes to set up a VDS. The acoustic nonce embedding and decoding process occurs while user is interacting with VDS. The computational overhead is negligible because the nonce size is small, and the playback device is used only when the user is speaking.

Reliability Next, we measure the reliability of embedding and retrieving acoustic nonce in the three environments. For this, we compute the average BER over 5 independent trials at different distances between the VDS input device (microphone) and user. Figure 8 shows that in all three environments, the average BER upto a distance of 1.5 m is 0%. Note, however that BER increases significantly beyond 1.5m in dining hall because the environment is relatively noisy.

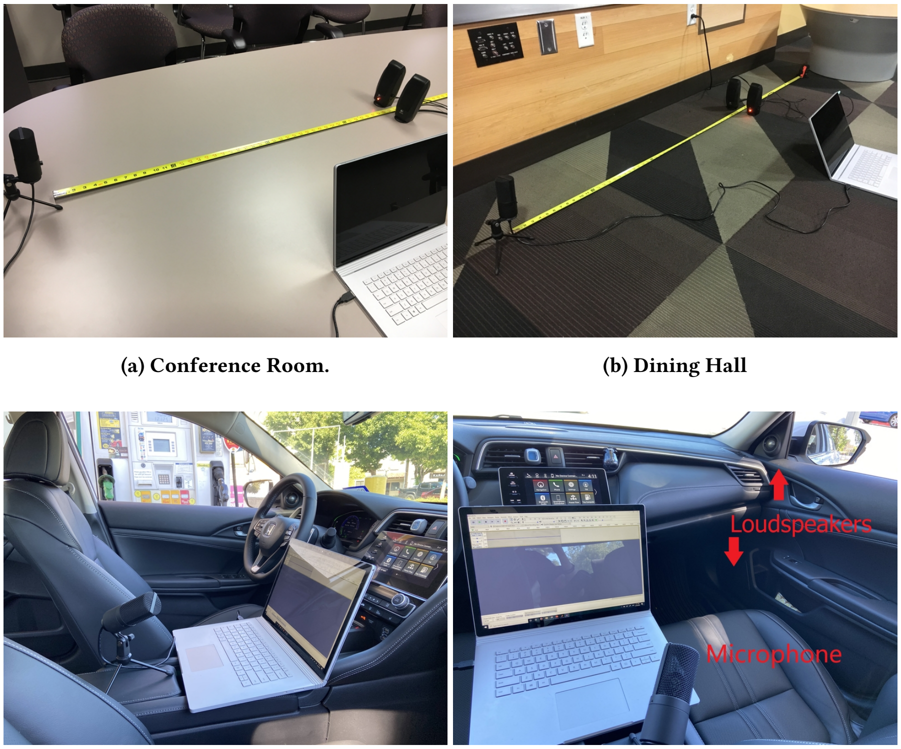
Speech Re-use Detection. In this experiment, we measure the efficacy of AEOLUS in detecting speech re-use (see Table 2). The input speech is considered re-used if (i) the current nonce is not decoded correctly, and (ii) if any trace of an obsolete nonce is detected. FRR indicates the proportion of falsely rejected samples because of incorrect nonce decoding or because the detected nonce is obsolete. FAR indicates how many re-used speech samples with obsolete nonces are falsely accepted. AEOLUS achieves 0.5% FRR at 0% FAR for speech re-use prevention upto a distance of 4 meters in the three environments.

5.2 Robustness

Impact on Speaker and Speech Recognition. The goal of this experiment is to assess the impact of acoustic nonce embedding and decoding on the performance of a commercial speaker and speech recognition system (Microsoft Azure). The results show that there is limited impact on the performance of the speaker recognition system for the conference room and dining hall environments upto a distance of 1.5 m (see Figure 9). The embedded nonce marginally increases the average word error rate (WER) of the speech recognition system to 4.4%. The WER is close to the reported average WER of 5.1% for the same system in the NIST 2000 Switchboard test set [60]. Manually reviewing the results indicates that most errors are related to incorrect suffix "-ing". A few errors are due to small word omissions such as "an" and "a". Relatively more errors occur in the dining hall environment due to the background noise.

Acoustic Nonce Removal. An adversary may attempt to remove the embedded acoustic nonce and obtain a "clean" speech sample to conduct re-use attacks. To test the robustness of AEOLUS in this adversarial setting, we use the following 6 common audio watermark removal techniques [59]:

- Resampling. Samples the audio at a different frequency (e.g. 44.1 KHz) to remove sampling-dependent watermark.
- Amplitude Compression. Alters the amplitude of recorded audio to bypass amplitude-related watermark detection.
- Filtering. Uses a high pass or low pass filter to remove a specific range of frequency in the audio. For example, separate audible signals from inaudible frequencies (e.g. ultrasound).
- Lossy Compression. Leverages data encoding methods that uses inexact approximations and partial data discarding to compress the audio. For example, MP3 [10] or MPEG-1 Audio Layer III reduce audio file size by taking advantage of a perceptual limitation of human hearing.



(a) Conference Room. **(b) Dining Hall**

(c) Gas Station (Car). **(d) Setup in the Car.**

Figure 7: Experimental setup to evaluate AEOLUS in different environments.

Table 1: Optimal parameters for achieving reliability and imperceptibility at a distance of 0.5 m between the microphone and user in different environments: (a) conference room, (b) dining hall, and (c) in a car parked at a gas station.

Location	SPL (dB-A)	Frequency (± 200 Hz)	Amplitude (normalized)	Bitrate (bits/sec)
Conference Room	41	4000 Hz	0.52	35
Dining Hall	58	5200 Hz	1	25
Gas Station (Car)	47	4800 Hz	0.61	33

Table 2: Acoustic nonce recovery bit error rate (BER) and speech re-use detection performance (FAR and FRR) at different distances in the three environments.

Location	Dis.	BER/FAR/FRR	Location	Dis.	BER/FAR/FRR	Location	Dis.	BER/FAR/FRR
Conference Room	0.5m	0%/0%/0%	Dining Hall	0.5m	0.2%/0%/0.5%	Gas Station (Car)	0.5m	0%/0%/0%
	1m	0%/0%/0%		1m	1.2%/0%/ 1%		0.7m	0.5%0%0.5%
	4m	1.2%/0%/0.5%		2m	75%/0% /78%			

- Equalization. Modifies the frequency response of an audio system using linear filters. Frequency response is the measure of the audio output in comparison to the input as a function of frequency. For example, a 5kHz wave input may have a 4kHz output response.
- Additive Noise. Adds noise to decrease SNR (signal-to-noise ratio) so that watermark cannot be extracted successfully.

Each technique is iteratively used until the acoustic nonce cannot be detected in a speech sample. The removal attempts are repeated for all speech samples in the dataset mentioned above. Table 3

Table 3: Performance of Microsoft Azure speaker recognition system (similarity score 0 - 1) and speech recognition system (WER) after nonce removal using two different FHSS schemes. Without nonce removal, the respective average similarity score and WER of the two systems are 0.897 and 5.1% [60]. Lossy compression and equalization nonce removal techniques failed to remove the nonce and the speech samples were rejected by AEOLUS.

Removal Technique	4-5kHz, 2 Hop., Avg. of 30 samples		1-8kHz, 10 Hop., Avg. of 30 samples	
	Speaker Verification (Score, 0-1)	Speech Recognition (WER %)	Speaker Verification (Score, 0-1)	Speech Recognition (WER %)
Resampling	0.233	29.73%	0.170	91.83%
Amplitude Compression	0.217	34.31%	0.184	63.4%
Filtering	0.623	17.12%	0.319	62.21%
Additive Noise	0.318	53.13%	0.192	96.34%
Lossy Compression	Reject	Reject	Reject	Reject
Equalization	Reject	Reject	Reject	Reject

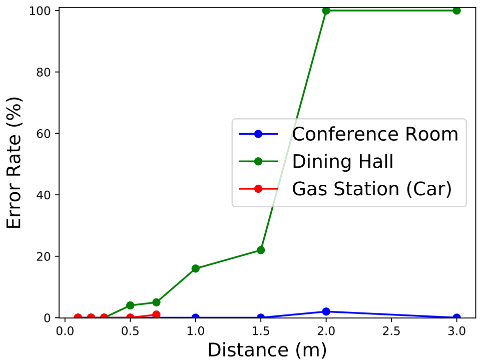


Figure 8: Variation of bit error rate (%) with different distances between microphone and user (simulated using a playback device) in different environments: (a) conference room, (b) dining hall, and (c) in a car parked at a gas station. Given a fixed size car in (c), results can only be computed upto a distance of 0.7 m.

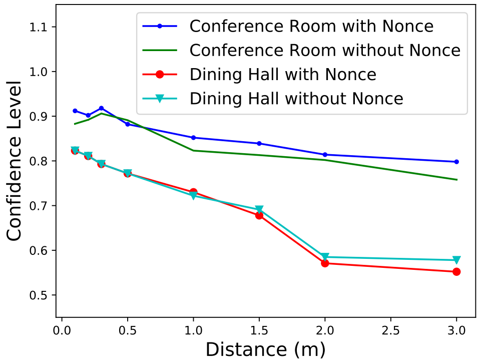


Figure 9: Impact of the proposed security overlay on prediction confidence of Microsoft Azure Speaker Recognition system in three different environments.

shows that nonce removal techniques are disruptive and ineffective in removing the acoustic nonce. This is because they reduced the average recognition performance of the commercial speaker recognition system by 79.52%, and increased the average word error rate of the speech recognition system by 33.57%. This shows that AEOLUS is robust against these removal techniques.

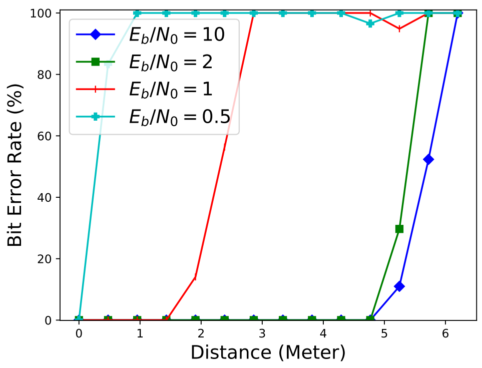


Figure 10: Impact of distance between the VDS loudspeaker and an adversary's recording device at the time of capture on speech re-use.

5.3 Distance between adversary and VDS speaker.

To launch a speech re-use attack, an adversary first records a user's interaction with VDS. In this experiment, we measure how effectively the acoustic nonce is preserved with varying distance between the VDS speaker and adversary's recording device (microphone) at the time of capture. Figure 10 shows that this distance significantly impacts BER. The computed BER is 0% upto a distance of 4 m when the signal to noise ratio is low. If the captured speech is being re-used, assuming a lossless audio input to VDS, the entire nonce can be recovered and used for determining the particular user interaction session the speech was previously used in.

5.4 Human Perception

5.4.1 Empirical Measurement. In this experiment, we empirically measure the change in SPL due to the embedded nonce. The distance between VDS output and the recording device is 1 m in dining hall and conference room, and 0.7 m in the car. The results (see Figure 11)

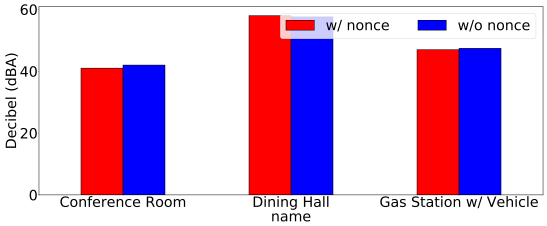


Figure 11: Loudness of speech with and without embedded nonce (average value of 5 measurements).

show that the average SPL difference in each environment is less than 1.4%. Non-parametric Wilcoxon Signed Rank test [58] (at 95% confidence level) is used to measure the statistical significance of the difference in SPL before and after acoustic nonce embedding. The results indicate that the differences are statistically insignificant.

User Study. Next, we study the impact of AEOLUS on VDS usability. For this, 120 participants (with prior experience using a VDS) from various demographics are recruited on Amazon Mechanical Turk. The demographics for the participants are shown in Table 6. There are more female participants than male participants, and 46.7% of the participants are between age 25 to 35. Note that this user study has been approved by the Institutional Review Board (IRB). The participants were given the option to opt-out at any time. No personally identifiable information is collected. All data were de-identified and stored securely.

The survey consists of three parts: device adjustment, preliminary questions, and survey questions. At the beginning of each survey, we guided the participants to adjust the loudness of their playback devices to a comfortable level. This is because the participants are recruited online, it is difficult to explicitly know or enforce the setup the participants use. Also, we pre-processed the recorded audio samples to ensure all audio samples were played with reasonable and consistent loudness.

Each participant is then asked to enter their demographic information. The participant is then instructed to play a test sample and adjust the playback volume to a comfortable level. Following this, the participant is presented with a survey with five speech samples and follow up questions. Included in the five samples is one sample with only background noise (and no speech) to check if the participant is attentive. The participants are asked to provide speech transcriptions for all samples and the provided transcriptions are used to check the participants' attentiveness. Eleven participants failed to pass this test and hence were excluded. Thus, there are 109 valid participants in this study (n = 109).

Next, each participant is presented with a survey that takes approximately 10 minutes to complete. Post completion of the survey, a participant is paid 4 USD as incentive. The survey questions are designed to ensure bias is not introduced in the results (no leading questions). We showcase the survey questions for each audio in Table 5. When designing the questions, we do not ask participants directly about the imperceptibility. Instead, we ask them whether they can transcribe the speech command in the audio. Second, we ask how noisy they find the audio (from 0-10, and 10 is the noisiest score). Third, we ask if any available noise affects the user experience. Fourth, if the noise affects their user experience, what does the noise sound like. This question is designed to verify whether it is the acoustic nonce that bothers them.

The test samples are uniformly selected at random from a dataset of 24 speech recordings, 12 recordings each with and without embedded nonce from the three environments. The average length of the speech samples is 6.4 seconds. To ensure the user study is unbiased, all samples presented to a participant are from the same speaker and pertain to the same voice command. After listening to a speech sample, a participant is asked to rate the noise level of the speech sample on a scale of 0 - 10 (10 being the highest noise level). The participant is also instructed to report any noticeable noise source that affects their experience, and to specify the characteristic of the noise that resulted in this disruption (e.g., human conversation, train or car noise). The answer is used to ascertain the source of noise that impacted the usability (e.g., the embedded nonce).

Results. Table 4 reports the average human perception ratings. The acoustic nonce does not degrade overall user experience for 94.16% of speech samples, on average, in the three environments. Since this does not adequately measure the participant's perception of nonce embedding for each sample, per speech sample perception is also reported. No speech sample with embedded nonce is perceived to be disruptive by the majority of participants.

We show some example comments from participants who reported that the noise in the audio affects their user experience,

> *Compared to the other samples, the background voices are much louder. In fact, a man speaking almost overpowered the voice I is supposed to listen to.* (P-004, G2[2])

We find some participants reported other noise sources but the acoustic nonce. For example, P-004 says the people chatting in the dining hall affected his user experience of VDS.

> *Car noises, but the singing is also very far away and hard to hear.* (P-072, G3)

Similarly, P-072 is complaining about street noise in Q5.

> *An echo, or like it is far away.* (P-019, G1)

Only 2 audios were reported to contain acoustic nonce (in Q5). For example, P-019 reported the echo sound and claimed it is affecting her user experience of VDS. However, we notice that the acoustic nonce didn't affect their transcriptions' correctness (i.e., Q2 in Table 5).

Perceived Noise Level. Table 7 shows that the speech samples containing the acoustic nonce are perceived to have a higher noise level. The dependent samples paired t-test (at 95% confidence) indicates a statistically significant difference (between samples with and without nonce) in the conference room and car. However, this difference is statistically insignificant in the dining hall due to higher background noise level compared to the conference room and car. Despite perceiving higher noise level, the majority of participants do not perceive any speech sample to be disruptive (see Table 4).

[2] G1, G2, G3 are respectively the participant groups for testing audios collected in the conference room, dining hall, and gas station.

Table 4: Perception of speech samples. Overall (%) is the proportion of participants that label speech samples as imperceptible, non-disruptive, or disruptive. Note that not all 109 valid participants listened to each location's recording (as mentioned above, random speech samples are played for each survey, and 11 participants are excluded). Sample (%) is the proportion of speech samples (with nonce) that are determined to be imperceptible, non-disruptive, or disruptive by the majority of participants.

Location	Imperceptible (Overall/Sample%)	Non-disruptive (Overall/Sample%)	Disruptive, Nonce-caused (Overall/Sample%)
Conference Room	48.1%/50%	50%/50%	1.9%/0%
Dining Hall	47.6%/25%	44.9%/75%	7.5%/0%
Gas Station (Car)	47.7%/25%	44%/75%	8.3%/0%

Table 5: User Study Survey Questions.

Question	Options
Q1. Pretend you are using voice-based service in a dining area, can you hear the speech in this audio clearly?	A. Very Clear B. Clear. C. It is neither clear or unclear but I am fine with it. D. I can hear the speech but it is very unclear. E. I cannot hear any speech.
Q2. Please transcribe the speech.	[Text]: ____
Q3. How noisy is this audio, rate from 0 - 10? (10 is most noisy)	[0-10]: ____
Q4. Do you find any noise affects your user experience?	A. I can hear something, but nothing bothers me. B. Yes, there is something troubles me. C. Nothing at all
Q5. If there are some noises affect your user experience for the voice service, what is it?	[Text]: ____

Table 6: User study participant demographic information (for all 120 participants).

Category	Gender Groups			Age Groups			
	Male	Female	Others	18-25	25-35	35-45	45-55
Number	53	65	2	32	56	21	11

6 DISCUSSION

In this section, we discuss AEOLUS's susceptibility to Denail of Service (DoS) attacks, its potential use in a layered VDS security solution and as a security indicator, and future work.

Denial of Service (DoS) Attacks. An adversary can launch a type of DoS attack called *signal jamming* on AEOLUS-enabled VDS by injecting acoustic waves with characteristic frequencies and amplitudes similar to AEOLUS's. However, this type of attack can be countered by frequently changing AEOLUS's operating frequency range. This would make it difficult for an attacker to predict the operating frequency at a given time and consequently launch effective real-time signal jamming attack. A smarter adversary can potentially inject acoustic waves in the entire spectrum of audible frequencies and amplitudes. Unlike common DoS attacks (e.g., link-flooding attacks [37, 42] in the network security domain) where it is usually challenging to pinpoint the attack, such an attack can be easily detected and mitigated. This is because to conduct the attack, the adversary would use the same frequency range as AEOLUS's (i.e., human audible frequencies), and these frequencies can be captured and analyzed without requiring any extra hardware.

Multiple AEOLUS-enabled VDS devices used around the same time in each other's neighborhood may accidentally result in a DoS scenario. A potential way to prevent this could be through a negotiable frequency range assignment (either centralized or decentralized) for different VDS devices to ensure they use different frequency ranges and nonce. Another reactive prevention method could be to allow AEOLUS to change its operating frequency range if a DoS situation is detected.

Layered VDS Security. AEOLUS is designed to prevent speech reuse as a security overlay. When used in conjunction with other security mechanisms, AEOLUS can greatly enhance VDS usability as a passive and near-imperceptible defense mechanism. Consider, for example, active liveness detection methods such as audio CAPTCHA [50] or challenge-response methods which require users to respond with dynamic words or passphrases. Although such methods provide additional security, they add friction from a usability standpoint, especially for short user interaction sessions. A potential risk-based layered approach could proactively use the proposed framework in every user interaction, and invoke challenge-response, to mitigate transaction risk as necessary.

Use as security indicator. The proposed framework is designed to embed near-imperceptible acoustic nonce in the human audible frequency range. However, it can be tailored for use as a slightly perceptible security indicator similar to electric vehicle warning

Table 7: Average perceived noise (on a scale of 0-10) in speech samples with and without acoustic nonce collected in the three environments.

Location	w/o Nonce	w/ Nonce	t-test, $\alpha = 0.05$	Statistically Significant
Conference Room	3.8	4.8	$t = 4.2022, p = 0.0001489$	Yes
Dining Hall	5.7	6.1	$t = 1.4595, p = 0.1524$	No
Gas Station (Car)	5	5.6	$t = 3.0697, p = 0.00389$	Yes

sound [46] or the light ring [17, 33] used in Amazon Echo devices. Presence of a gentle perceptible noise at the time of user interaction could indicate to the user that the interaction session with VDS (e.g., Amazon Alexa) is secured using AEOLUS.

Future Work. In future, we will conduct in-depth investigation of methods for generation and encoding of acoustic nonce, e.g., using variations of music or other forms of sound along with different encoding schemes. We also plan to conduct large-scale live subject testing to further evaluate reliability and imperceptibility. To address the scenario when clean speech can be obtained by an adversary, we will investigate both passive strategies, e.g., use of acoustic environment identification techniques to detect a voice command recorded in a different environment, and active strategies, e.g., prompting users to speak a session-specific random number in voice commands.

7 SUMMARY

In this work, we present a security overlay called AEOLUS that can be integrated with any voice-driven services to prevent speech re-use. AEOLUS reliably embeds a dynamic acoustic nonce that is non-disruptive to a VDS user, and detects the embedded nonce from the recorded user speech. Experiments conducted in three different environments show that AEOLUS can be used in operational VDS scenarios with minimal overhead.

ACKNOWLEDGMENTS

This material is based upon work supported in part by the NSF/VMware Partnership on Software Defined Infrastructure as a Foundation for Clean-Slate Computing Security (SDI-CSCS) program under Award Title "S2OS: Enabling Infrastructure-Wide Programmable Security with SDI" and No. 1700544. It is also supported in part by ONR Grant No. N00014-20-1-2734. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF and ONR.

REFERENCES

[1] 2020. Alexa for Business. https://aws.amazon.com/alexaforbusiness/.
[2] 2020. Amazon Alexa Pay in Gas Station. https://bit.ly/37v25xN.
[3] 2020. Black Hat 2018: Voice Authentication is Broken. https://bit.ly/3mxQIto.
[4] 2020. Equal-loudness contour. https://bit.ly/2LF1ukH.
[5] 2020. Face ID Security. https://apple.co/34mmO59.
[6] 2020. IoSR Matlab Toolbox: Equal-loudness Contour Mapping. https://bit.ly/2Wtd1Wz.
[7] 2020. Just say it: The future of search is voice and personal digital assistants. https://www.campaignlive.co.uk/article/just-say-it-future-search-voice-personal-digital-assistants/1392459.
[8] 2020. Microsoft Azure Speaker Recognition - Verification. https://bit.ly/3h82Sbh.
[9] 2020. Microsoft Cognitive Services. https://bit.ly/38haWCH.
[10] 2020. MP3 Wikipedia. https://en.wikipedia.org/wiki/MP3.
[11] 2020. Paysafe Study. https://bit.ly/2WtgjJl.
[12] 2020. Personal Music Players & Hearing. https://bit.ly/34tyWkN.
[13] 2020. Voice Password in Wechat. https://zd.net/3myAGQ2.
[14] 2020. Voice Payment Survey. https://bit.ly/3p6k97z.
[15] 2020. Will voice assistant change workplace. https://bit.ly/3mxRDKm.
[16] 2021. Alexa skill: 300,000 daily users. https://vux.world/google-assistant-500m-active-users-popular-alexa-skill-300000-daily-users/.
[17] 2021. Echo Device Light Indicator. https://www.amazon.com/gp/help/customer/display.html?nodeId=GKLDRFT7FP4FZE56.
[18] 2021. Paypal Voice Assistant Application with Pins. https://bit.ly/3h8QkR6.
[19] 2021. Voice Password Applications in Google Play. https://bit.ly/38dq1oG.
[20] 2021. Voice Password in Samsung Bixby. https://bit.ly/34rlPAN.
[21] H. Abdullah, W. Garcia, C. Peeters, P. Traynor, K.R.B. Butler, and J. Wilson. 2019. Practical hidden voice attacks against speech and speaker recognition systems. *arXiv preprint 1904.05734* (2019).
[22] M. E. Ahmed, I. Kwak, J. H. Huh, I. Kim, T. Oh, and H. Kim. 2020. Void: A fast and light voice liveness detection system. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*. 2685–2702.
[23] P. Bassia, I. Pitas, and N. Nikolaidis. 2001. Robust audio watermarking in the time domain. *IEEE Transactions on multimedia* 3, 2 (2001), 232–241.
[24] M. Bellare and P. Rogaway. 1995. Provably secure session key distribution: the three party case. In *ACM symposium on Theory of computing*.
[25] G. Chen, S. Chen, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu. 2021. Who is real bob? adversarial attacks on speaker recognition systems. *IEEE Symposium on Security and Privacy* (2021).
[26] S. Chen, K. Ren, S. Piao, C. Wang, Q. Wang, J. Weng, L. Su, and A. Mohaisen. 2017. You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones. In *International Conference on Distributed Computing Systems*.
[27] T. Chen, L. Shangguan, Z. Li, and K. Jamieson. 2020. Metamorph: Injecting inaudible commands into over-the-air voice controlled systems. *Proceedings of NDSS* (2020).
[28] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon. 1996. A secure, robust watermark for multimedia. In *International Workshop on Information Hiding*.
[29] N. Cvejic. 2004. *Algorithms for audio watermarking and steganography.*
[30] N. Cvejic and T. Seppänen. 2004. Spread spectrum audio watermarking using frequency hopping and attack characterization. *Signal processing* 84, 1 (2004), 207–213.
[31] S. Das and P. N. Suganthan. 2010. Differential evolution: A survey of the state-of-the-art. *IEEE transactions on evolutionary computation* 15, 1 (2010), 4–31.
[32] H. Feng, K. Fawaz, and K. G. Shin. 2017. Continuous authentication for voice assistants. In *International Conference on Mobile Computing and Networking.*
[33] C. Geeng. 2020. EGregor: An Eldritch Privacy Mental Model for Smart Assistants. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.
[34] Y. Gong and C. Poellabauer. 2018. Protecting voice controlled systems using sound source identification based on acoustic cues. In *International Conference on Computer Communication and Networks.*
[35] Y. Gong, J. Yang, J. Huber, M. MacKnight, and C. Poellabauer. 2019. ReMASC: Realistic Replay Attack Corpus for Voice Controlled Systems. *arXiv preprint 1904.03365* (2019).
[36] D. H. Johnson. 2006. Signal-to-noise ratio. *Scholarpedia* 1, 12 (2006), 2088.
[37] Min Suk Kang, Virgil D Gligor, Vyas Sekar, et al. 2016. SPIFFY: Inducing Cost-Detectability Tradeoffs for Persistent Link-Flooding Attacks.. In *NDSS*, Vol. 1. 53–55.
[38] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee. 2017. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. (2017).
[39] D. Kirovski and H. S. Malvar. 2003. Spread-spectrum watermarking of audio signals. *IEEE transactions on signal processing* 51, 4 (2003), 1020–1033.
[40] P. Kumar, M. Jayakumar, and A.V. Vidyapeetham. 2010. Comparison of bit error rate for propagation mechanisms of millimeter waves in a practical communication systems employing PSK and FSK. *PIERS Proceedings* (2010).
[41] E. A. Lee and D. G. Messerschmitt. 2012. *Digital communication.* Springer Science & Business Media.
[42] S. B. Lee, M. S. Kang, and V. D. Gligor. 2013. CoDef: Collaborative defense against large-scale link-flooding attacks. In *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies.* 417–428.

[43] M. Leng, S. S. Arora, and K. Wagner. 2020. Replay spoofing detection for automatic speaker verification system. US Patent App. 16/535,836.

[44] C. Neubauer and J. Herre. 1998. Digital watermarking and its influence on audio quality. In *Audio Engineering Society Convention 105*. Audio Engineering Society.

[45] J. Park, K. Park, and J. R. Yoon. 2010. Underwater acoustic communication channel simulator for flat fading. *Japanese Journal of Applied Physics* 49, 7S (2010).

[46] P. Poveda-Martinez, R. Peral-Orts, N. Campillo-Davo, J. Nescolarde-Selva, M. Lloret-Climent, and J. Ramis-Soriano. 2017. Study of the effectiveness of electric vehicle warning sounds depending on the urban environment. *Applied Acoustics* 116 (2017), 317–328.

[47] S. Pradhan, W. Sun, G. Baig, and L. Qiu. 2019. Combating Replay Attacks Against Voice Assistants. *ACM Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019).

[48] N. Scaife, C. Peeters, and P. Traynor. 2018. Fear the reaper: Characterization and fast detection of card skimmers. In *27th USENIX Security Symposium*.

[49] J. W. Seok and J. W. Hong. 2001. Audio watermarking for copyright protection of digital audio data. *Electronics Letters* 37, 1 (2001), 60–61.

[50] Y. Soupionis and D. Gritzalis. 2010. Audio CAPTCHA: Existing solutions assessment and a new implementation for VoIP telephony. *Computers & Security* 29, 5 (2010), 603–618.

[51] R. Storn and K. Price. 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization* 11, 4 (1997), 341–359.

[52] Y. Suzuki and H. Takeshima. 2004. Equal-loudness-level contours for pure tones. *The Journal of the Acoustical Society of America* 116, 2 (2004), 918–933.

[53] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney. 1998. Robust audio watermarking using perceptual masking. *Signal processing* 66, 3 (1998), 337–355.

[54] B. Truax. 2001. *Acoustic communication.* Greenwood Publishing Group.

[55] J. L. Tsai. 2009. Efficient Nonce-based Authentication Scheme for Session Initiation Protocol. *IJ Network Security* 9, 1 (2009), 12–16.

[56] A. Wabnitz, N. Epain, C. Jin, and A. Van Schaik. 2010. Room acoustics simulation for multichannel microphone arrays. In *International Symposium on Room Acoustics*.

[57] S. Wang, J. Cao, X. He, K. Sun, and Q. Li. 2020. When the Differences in Frequency Domain are Compensated: Understanding and Defeating Modulated Replay Attacks on Automatic Speech Recognition. In *ACM SIGSAC Conference on Computer and Communications Security*.

[58] F. Wilcoxon, S. K. Katti, and R. A. Wilcox. 1970. Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. *Selected tables in mathematical statistics* 1 (1970), 171–259.

[59] S. Wu, J. Huang, D. Huang, and Y. Q. Shi. 2005. Efficiently self-synchronized audio watermarking for assured audio data transmission. *IEEE Transactions on Broadcasting* 51, 1 (2005), 69–76.

[60] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke. 2018. The Microsoft 2017 conversational speech recognition system. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

[61] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. Dolphinattack: Inaudible voice commands. In *ACM SIGSAC Conference on Computer and Communications Security*.