# Design of Multi-Sine Watermark using Power Spectral Analysis for Replay Attack Detection

**Sunitha George**

**A Thesis in**

**The Department**

**of**

**Electrical & Computer Engineering**

**Presented in Partial Fulfilment of the**

**Requirements for the Degree of**

**Master of Applied Science (Electrical & Computer Engineering)**

**at Concordia University**

**Montreal, Quebec, Canada**

**June 2025**

# CONCORDIA UNIVERSITY

## School of Graduate Studies

This is to certify that the thesis prepared

By: Sunitha George

Entitled: Design of Multi-Sine Watermark using Power Spectral Analysis for Replay Attack Detection

and submitted in partial fulfilment of the requirements for the degree of

**Master of Applied Science (Electrical & Computer Engineering)**

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair

*Dr. R. Soleymani*

_____ External Examiner

*Dr. F. Haghighat*

_____ Examiner

*Dr. R. Soleymani*

_____ Thesis Supervisor

*Dr. Shahin Hashtrudi Zad*

Approved by _____

*Dr. R. Soleymani, Chair*

*Department of Electrical & Computer Engineering*

12 June 2025 _____

*Mourad Debbabi, Dean*

*Faculty of Engineering and Computer Science*

# Abstract

Design of multi-sine watermark using power spectral analysis for replay attack detection

Sunitha George

Replay attacks are a critical security concern in cyber-physical systems (CPS), where adversaries record legitimate data transmissions and maliciously retransmit them later to disrupt normal system operations. These attacks are particularly dangerous because they often replay legitimate data, making them difficult to detect using traditional intrusion detection systems. As CPS continue to integrate deeper into critical infrastructure such as power systems, industrial automation, and transportation networks, the need for better safety measures becomes increasingly urgent.

One promising line of defense involves watermarking techniques, in particular, using multi-sine watermarks with switching frequencies. This thesis studies the problem of choosing the parameters of multi-sine watermarks to achieve replay attack detection with desired level of confidence. The proposed method is derived from a power spectral analysis of the output of the plant in both normal (no attack) and during attack operation.

A flow control process involving a tank is utilized as an illustrative example. Through this example, the effectiveness of the proposed method is validated, showing its capability to design a watermark that can successfully detect replay attacks and thus enhance the security of the control system.

# Acknowledgement

First and foremost, I give all praise and thanks to God Almighty, whose grace and guidance have been my constant companions. As it is written in Philippians 4:13, "***I can do all things through Christ who strengthens me.***" This achievement would not have been possible without His divine presence, which has given me strength in moments of doubt and perseverance through every challenge.

I would also like to extend my sincere gratitude to my supervisor, Dr. Shahin Hashtrudi Zad. Your invaluable guidance, patience, and expertise have been instrumental in shaping this work. Thank you for believing in my potential and pushing me to reach heights I didn't know I was capable of. Your support and constructive feedback have made this thesis a rewarding journey, and for that, I am forever grateful.

To my beloved parents and brother, words cannot fully express my gratitude for your unwavering love, encouragement, and support. You have been my foundation, guiding me with your wisdom, and lifting me up with your faith in my abilities. Your sacrifices, belief, and tireless motivation have been my driving force, and I am deeply grateful for every step you've walked alongside me on this journey.

To my fiancé, thank you for your endless patience, love, and steadfast encouragement. Your understanding and support have given me the strength to push forward, and your presence has been a constant source of comfort during this process. I am truly blessed to have you by my side.

This accomplishment is the result of the collective love, strength, and guidance that each of you has so generously provided. I am deeply thankful to God and to all of you for being the pillars of my success.

# Table of Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| **ICS** | Industrial Control Systems |
| **CPS** | Cyber-Physical System |
| **DoS** | Denial of Service |
| **GRA** | Generalized Replay Attack |
| **GMC** | Generic Model Control |
| **LQZ** | Linear Quadratic Zonotopic |
| **DOF** | Degree of Freedom |
| **PSD** | Power Spectral Density |
| **IoT** | Internet of Things |
| **SNR** | Signal to Noise Ratio |
| **WSS** | Wide Sense Stationary |
| **BPF** | Band Pass Filter |
| **SCADA** | Supervisory Control and Data Acquisition |
| **FDI** | False Data Injection |
| **LQG** | Linear Quadratic Gaussian |
| **AGMC** | Adaptive Generic Model Control |
| **NCS** | Networked Control System |
| **LTV** | Linear Time-Varying |
| **IID** | Independent and Identically Distributed |

| | |
|---|---|
| **DFT** | Discrete Fourier Transform |
| **AGC** | Automatic Generation Control |
| **FFT** | Fast Fourier Transform |
| **LTI** | Linear Time Invariant |

# Chapter 1

# Introduction

A cyber-attack on control systems involves malicious activities aimed at disrupting, damaging, or gaining unauthorized control over Industrial Control Systems (ICS) and Supervisory Control and Data Acquisition (SCADA) systems, which are widely used in critical infrastructure sectors such as energy, water treatment, transportation, manufacturing, and utilities. These systems monitor and control industrial processes, ensuring the safe and efficient functioning of physical machinery. When compromised, the impacts can range from operational downtime and financial loss to severe physical damage, environmental hazards, or even threats to human life [1]-[2].

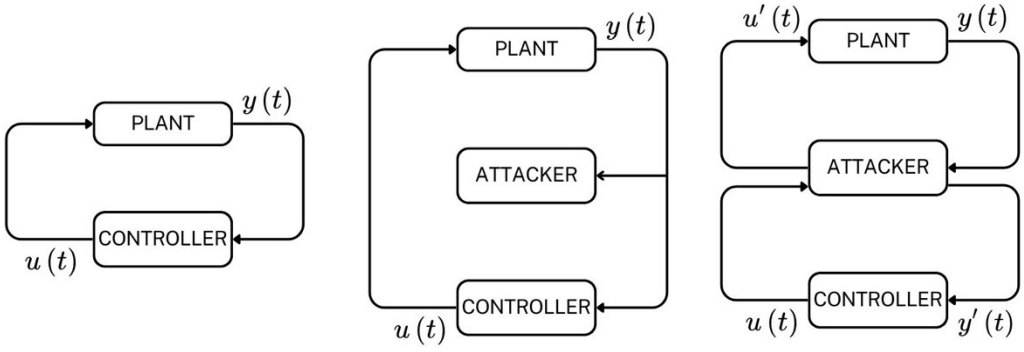
In this thesis, the problem of detecting one group of cyber-attacks know as replay attacks is examined. Replay attacks are a type of cyber threat in which previously recorded legitimate data is captured and retransmitted to the system at a later time, with the goal of misleading the system into accepting false information as genuine. These attacks are particularly difficult to detect because the replayed data closely mimics normal system behaviour, often bypassing standard security measures. As Cyber-Physical Systems (CPS) become increasingly integrated into critical infrastructure, the risks associated with such attacks are heightened [3]-[4]. Therefore, the detection of replay attacks is treated as a critical security concern.

## 1.1 Cyber Attacks on Control Systems

Cyber-attacks on control systems have become an escalating concern with the increasing digitization and connectivity of industrial infrastructure. These systems, which govern critical processes in the energy, manufacturing, and transportation sectors, have been targeted by a range of sophisticated cyber threats. Notable real-world incidents include the Stuxnet worm [5], which in 2010 was used to sabotage Iran's nuclear centrifuges by manipulating control system behaviour while hiding its presence, and the BlackEnergy malware [6], which played a key role in the 2015 Ukrainian power grid attack that caused widespread outages. Another significant incident involved the Triton malware [7], which targeted safety instrumented systems in a Saudi petrochemical plant, potentially putting human lives at risk. In addition to replay attacks, which involve the

retransmission of previously captured valid data to deceive systems, other common cyber threats include False Data Injection (FDI) attacks [8], where attackers corrupt sensor or control data to mislead system decisions, and denial-of-service (DoS) attacks [9], which aim to overwhelm system resources, causing loss of control or functionality. These cases highlight the vulnerability of control systems and emphasize the need for advanced detection and mitigation strategies. In this context, understanding and addressing specific types of attacks—such as replay attacks—is essential to safeguarding the stability and safety of modern control systems.

A replay attack on an ICS typically unfolds in two key stages: recording and replay. In the recording stage ($Fig.\,1.1(b)$), the attacker passively intercepts and records legitimate communication between sensors and the controller, capturing normal system behaviour without altering any data or triggering alarms. In the replay stage ($Fig.\,1.1(c)$), the attacker replaces real-time sensor data with the previously recorded sequence, effectively deceiving the controller into responding to outdated but valid data while the actual system state may drift into unsafe or undesired conditions. Because the replayed data was once legitimate, this type of attack can bypass many traditional detection mechanisms and pose a serious threat to the safety and reliability of industrial processes.




$(a)$ $(b)$ $(c)$

$Fig.\,1.1$. Block diagram of a control system $(a)$ under safe conditions $(b)$ being recorded by attacker and $(c)$ under a replay attack.

## 1.2 Literature Review

In this section, a review of existing research related to cyber-attacks on ICS, with a particular focus on replay attacks and the various strategies developed to detect them is presented. The discussion begins with an overview of common cyber threats targeting ICS, emphasizing the growing risk posed by stealthy attacks such as replay. It then explores the concept of replay attacks highlighting their unique challenges and the limitations of traditional detection techniques. Furthermore, the use of signal watermarking as a defence mechanism is examined, including different watermark designs and their integration into control systems. Recent advances in replay attack detection mechanisms and their practical applications are also surveyed, providing a foundation for the development of the methodology proposed in this thesis.

Cyber-attacks can have severe consequences, especially in ICS that manage critical infrastructure like power plants, water treatment facilities, and manufacturing processes [10]. If not mitigated, they can lead to operational disruption as attackers can cause devices or systems to repeat actions, leading to production line halts, equipment malfunctions, or process inefficiencies [11]. In safety-critical systems, cyber-attacks could trigger hazardous conditions, such as overloading equipment, causing leaks or shutdowns, or compromising safety controls [12]. Majority ICSs rely on old, unencrypted communication protocols that are vulnerable to interception and replay [13]. These systems lack advanced security features that can prevent replay attacks [14]. In some environments, security measures like encryption or message authentication are either absent or poorly implemented, making replay attacks even easier to carry out [15].

Watermarking in ICS is a security technique used to detect cyber-attacks, particularly replay attacks, by deliberately embedding known, distinguishable signals into the control loop. These signals, referred to as *watermarks*, are designed such that their presence can be verified in the sensor measurements, allowing the system to determine whether the data being received is genuine or has been tampered with or replayed [16]. Under normal operation, this watermark propagates through the plant and affects the sensor measurements. The detector can then check whether the expected effect of the watermark is present in the incoming data. However, during a replay attack, the attacker replays old sensor data that does not contain the current watermark signal. As a result,

the inconsistency between the applied watermark and the system response becomes detectable, indicating the presence of an attack. Watermarks are carefully designed to be:

- Undetectable to attackers, so they cannot be easily filtered out or replicated.
- Non-disruptive to system performance, ensuring control objectives are not compromised.
- Statistically or spectrally distinguishable, to aid in reliable detection.

In essence, watermarking converts the replay attack detection problem from a passive monitoring task to an active verification process, enhancing the resilience of ICS against stealthy adversaries [17].

Numerous attempts have been made in the past to design an effective watermarking technique with reliable detection methods. In [18], a design mechanism for injecting the watermarking signal using a plug and play approach without affecting the predefined system controller was proposed and illustrated using a numerical application example. The method of injecting a series of secret noisy control inputs called physical watermarking was explored in [18] where the adversary was assumed to have information access to the system. Model inversion physical watermarking was explored in [20] that simultaneously helped with performance tracking and security performance. An on-line approach to physical watermarking was presented in [21]. A physical dynamic watermarking design was implemented in [22] that was evaluated on a water distribution network.

A series of limitations and challenges were identified for physical watermarking. [23] analyzed the usage of finite sequences as physical watermarking method tested on a four-tank system, thereby detecting the replay attack using zonotopes. The use of zonotopes for the design of watermark signal was explored in [24] where the analogy between stochastic and deterministic fields was extended to optimal control of a Linear Quadratic Gaussian (LQG) system. A simultaneous synthesis of optimal watermarking signal and robust controller was proposed in [25] and implemented over a simplified three-tank chemical system.

Attempting to improve on the control effort, an Adaptive Generic Model Control is proposed and evaluated against the Generic Model Control [26]. The estimation of the time-varying parameters is made robust, and this method was found to perform well for the control of complex chemical processes. Attack models and scenarios for Networked Control Systems (NCSs) are a major focus

in the security of CPS [27]-[28], where control loops are closed through communication networks. NCSs are found in power systems, industrial control systems, autonomous vehicles, etc., making their resilience to cyber-attacks critically important. Control laws are designed to maintain stability and performance even under certain classes of attacks. Techniques such as watermarking, state estimation with anomaly detection or redundancy and secure communication protocols are generally used for safeguarding NCSs. A stochastic game approach was presented in [29] that aims to balance security overhead with control cost. Further, [30] proposed a countermeasure with a fixed false alarm rate to optimize the trade-off between detection delay and LQG performance by decreasing control accuracy or increasing control effort.

A lot of efforts to optimize the existing detection methods have also been made. Replay attack detection using a watermarked control strategy shared among the agents through the network was introduced in [31] with minimum loss effect. A strategy through sensor watermarking was also recorded [32]. A $\chi^2$ detector was modified by incorporating a watermarking signal through a random number generator in [33]-[34]. The modified $\chi^2$ detector was used to detect and mitigate FDI attack and replay attack. [35] proposed a replay attack detection method using a frequency-based signature method which used a sinusoidal signal with time-varying frequency as the watermark. The impact of a zonotopically bounded watermarking signal was assessed for replay attack detectability in [36] along with a Linear Quadratic Zonotopic (LQZ) controller. In [37], a Linear Time-Varying (LTV) dynamic watermarking was tested on a car model in CarSim and a 1/10 scale autonomous rover and was proven to quickly detect replay attacks in a repeatable fashion. Time-Varying Dynamic Watermarking was proposed to detect Generalized Replay Attacks (GRAs) [38] simulated using an LTV vehicle. Multiplicative watermarking based active detection of GRAs was detailed in [39]. A stealthy replay attack was detected using dynamic watermarking approach based modified $\chi^2$ detector [40] for a 3—DOF helicopter benchmark system. A dynamic watermarking method for detecting replay attacks was tested in controlled discrete LQG systems [41] and was found to have a better performance than IID watermarking signal. Optimal watermarking schemes in order to minimize the control costs are being researched upon on a large scale to facilitate detection of replay attacks better [42]-[43].

A switching multi-sine watermarking technique was introduced in [44]. The parameter choice for the technique was made with the main objective of transient response suppression. In [44], since

the energy of the watermarking signal was primarily concentrated at certain frequencies due to the nature of the sine signal, a periodogram was employed for detection purposes. The periodogram is a fundamental tool used to estimate the Power Spectral Density (PSD) of a signal, which represents how signal power is distributed across different frequencies. It involves computing the Discrete Fourier Transform (DFT) of a finite-length signal and then squaring the magnitude of the result, typically normalized by the signal length [45]. While simple and widely used, the basic periodogram can be a noisy estimate due to its high variance, especially for short data records. To improve reliability, techniques like windowing and averaging are often applied, making the periodogram a foundational tool in spectral analysis [46]. Precisely, the confidence bounds for periodogram were used to authenticate the watermarking signal by detecting the frequencies present in the signal.

Watermarking is relevant to other applications as well. One such example is that of smart grid systems. A random signal was added periodically for a small duration of time providing sufficient frequency detection capability proposing an effective detection method for replay attacks [47]. Another detection method using power state sampled to detect the replayed states is presented and evaluated on the IEEE bus systems [48]. In [49], a blockchain-based decentralized mechanism for replay attack detection is explored for large scale power systems. A comprehensive review on the countermeasures for replay attack on smart grids is presented in [50]. [51] employs the switching multi-sine watermarking technique that was introduced in [44] for replay attack detection in smart grids.

## 1.3 Thesis Objectives and Contributions

The main objective of the thesis is to develop and propose an effective method for designing multi-sine watermark signals for replay attack detection. The goal is to provide an algorithm for choosing the parameters of the multi-sine watermarking signal. This algorithm is intended to yield a multi-sine watermark custom made for a particular application authenticated using periodograms with a chosen level of confidence.

The major contributions of this thesis have been as follows.

1. An algorithm is developed for choosing watermark frequencies and watermark power to ensure a desired rate of false alarm.

2. In the process of the algorithm, an analysis of the power spectral density of watermarked output is performed that provides some of its characteristics. These results in future can be used to determine other performance measures such as detection time.
3. The proposed algorithm is used to design watermark for flow control system. The performance of watermarking is explored using computer simulation.

## 1.4 Thesis Outline

This thesis is structured into five chapters, each addressing key concepts related to watermarking and its analysis. Chapter 1 provided an overview of the research, beginning with a literature review that summarized relevant studies and contextualized the work within existing knowledge. It outlined the thesis objectives and contributions, specifying the research goals and unique contributions to the field. The thesis outline offers a structured overview of the subsequent chapters. Chapter 2 covers essential background material, introducing key concepts such as the replay attack followed by an exploration of multi-sine watermarking and a discussion on power spectral density using periodogram analysis. Chapter 3 focuses on multi-sine watermarking signal parameters, starting with the problem statement to identify challenges, then detailing the proposed design method, and concluding with a summary. Chapter 4 presents a case study, including the algorithm for using the proposed method and the simulation results obtained through testing, culminating in a conclusion that summarizes the findings. Finally, Chapter 5 wraps up the thesis by summarizing the key points and suggesting directions for future research to extend the work done in this study.

# Chapter 2

# Background

Control systems are particularly vulnerable to cyber-attacks due to their reliance on legacy technologies, limited patching capabilities, and increased connectivity with corporate IT networks and the internet. With growing digitization and the rise of the Internet of Things (IoT) in industrial environments, the attack surface is expanding, necessitating stronger cybersecurity measures to protect critical infrastructure from potentially catastrophic cyber incidents.

One of the techniques considered for securing control systems is the implementation of robust watermarking techniques, which serve as an effective means of detecting and mitigating cyber threats. Watermarking involves embedding a uniquely identifiable signal into system, allowing for the verification of system integrity and the detection of unauthorized intrusions. However, designing an effective watermarking scheme presents challenges, including maintaining a balance between detectability and stealth, minimizing the impact on system performance, and countering adversarial strategies aimed at removing or bypassing the watermark. Addressing these challenges requires a comprehensive understanding of control system, signal processing, and cybersecurity principles to develop resilient and adaptive defense mechanisms.

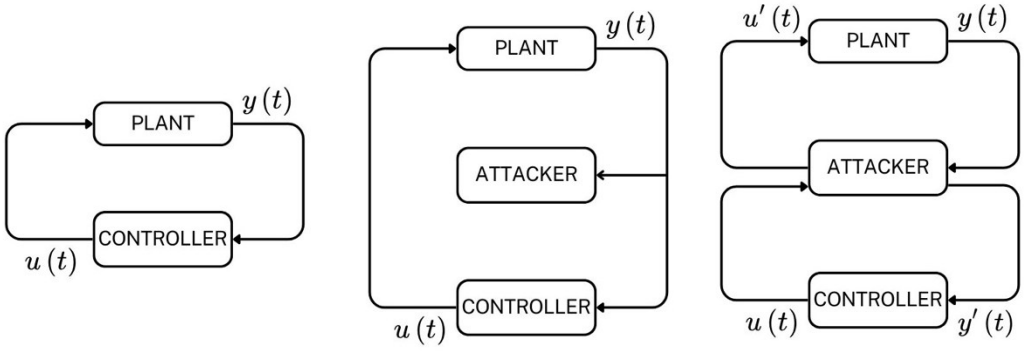
## 2.1 Replay Attack

A *replay attack* in control systems refers to a cybersecurity threat where an adversary intercepts, records, and later replays legitimate sensor readings to deceive the system into acting on outdated or manipulated data. This attack undermines system integrity by exploiting the system's inability to distinguish between real-time and previously captured signals, potentially leading to unsafe operational states or a loss of control and stability. Replay attacks are particularly harmful in critical infrastructures such as power grids, transportation and industrial automation where real-time data integrity is essential for safe and accurate functioning.

Replay attacks are relatively easy to perform because they do not require deep knowledge of the intercepted communication. The attacker does not need to understand the data. Simply capture and retransmit is sufficient. Many control systems and network protocols are designed with a focus on

efficiency and reliability rather than security. They often assume that once a message has been authenticated, it can be trusted without further checks. Replay attacks exploit this trust by using valid, previously authenticated data to bypass security controls.

A replay attack in control systems is particularly effective when launched after the system has reached steady state. In steady state, sensor readings and control signals exhibit minimal fluctuations. Since the replayed data is nearly identical to legitimate steady-state values, anomaly detection mechanisms may not flag the attack. When an actual disturbance occurs, the controller needs fresh sensor data to react properly. However, when old steady-state data is replayed, the controller remains unaware of the change, potentially leading to unsafe operation or instability.




$(a)$ $(b)$ $(c)$

$Fig.2.1$. System $(a)$ under safe conditions $(b)$ being recorded by attacker and $(c)$ under a replay attack [44].

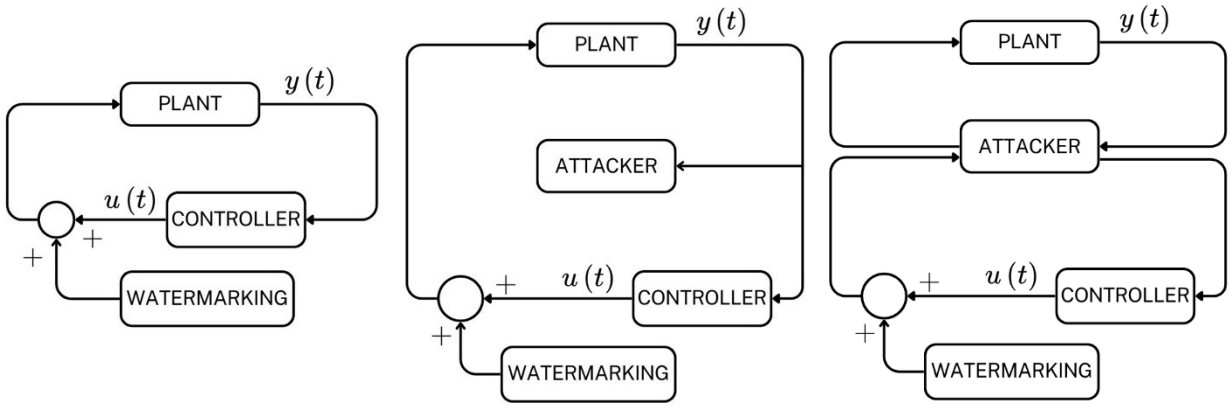
The attacker monitors the system communication until it reaches steady state and then records sensor data as shown in $Fig.2.1(b)$. Later during an attack window, the attacker replays the previously recorded data as shown in $Fig\ 2.1(c)$. This leads the controller to operate based on stale, replayed data rather than real-time conditions. The system fails to respond to disturbances or changes. In critical systems, this can cause physical damage or financial losses. For example, in a power grid, Automatic Generation Control (AGC) maintains frequency stability. If the attacker records frequency and power flow data when the system is stable and later, when demand fluctuates, if the attacker replays this steady-state data, then controller will not respond to real-time load changes, leading to grid instability or even blackouts. Furthermore, during a replay

attack, the attacker can replace control signal with another, creating disruption or damage to the plant ($Fig.\,2.1\ (c)$).

## 2.2 Replay Attack Detection using Watermarking

Watermarking is a technique widely applied in cybersecurity, particularly in image processing, to provide integrity and authenticity assurances by embedding identifying signals or patterns (the "watermark") within transmitted data. In the context of detecting replay attacks, watermarking serves as a proactive defense mechanism.

In control systems, watermarking typically involves adding a low-amplitude, pseudo-random, or deterministic signal to the input commands as shown in $Fig.\,2.2$. This embedded signal is chosen to be imperceptible in regular operation but detectable during verification processes. By altering the command signal with these watermarks, control systems can verify the temporal integrity of incoming data, ensuring that each received signal contains the watermark pattern.




$Fig.\,2.2$. Block diagram of system with watermarking $(a)$ under safe conditions $(b)$ being recorded by attacker and $(c)$ under a replay attack.

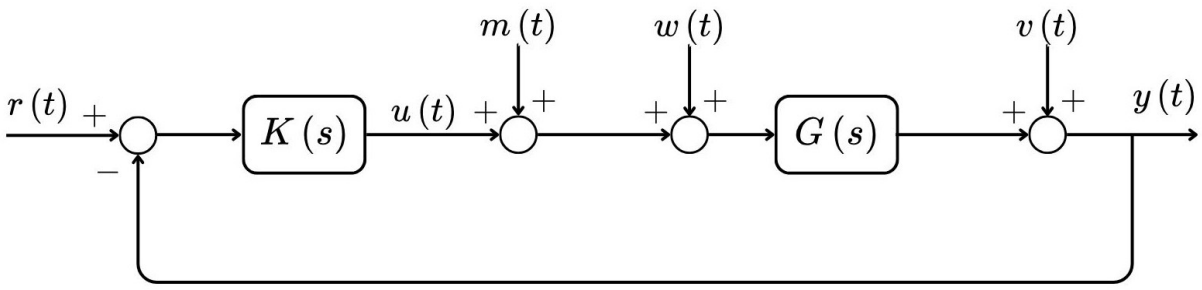
During a replay attack, when a replayed signal is introduced to a control system, the watermark patterns will misalign with expected values. As these watermarks are designed to change dynamically, the replayed data will likely lack the correct watermark sequence, making the replayed data immediately suspicious. The effects of watermarking in the system output during

attack is not consistent with the watermark at that time. This inconsistency can be used to detect the replay attack.

The watermark signal can be designed to be system-specific, leveraging parameters unique to the operational dynamics of the control system. This ensures that the watermark is neither easily replicated nor guessed by adversaries.

## 2.3 Multi-sine Watermarking

In this section, we review the switching multi-sine watermarking of [44]. Consider a single-input single-output plant under control at steady state around a set point. The block diagram of the linear model around the set point and watermarking is shown in $Fig.\,2.3$.




$Fig.\,2.3$. Block diagram of linear system with watermarking.

In $Fig.\,2.3$, $K(s)$ is the controller transfer function, $G(s)$ the plant transfer function, $r(t)$ the input signal, $y(t)$ the output sensor signal, $u(t)$ the control signal, $m(t)$ the watermarking signal, $w(t)$ the input disturbance and $v(t)$ the measurement noise.

Multi-sine watermarking uses sine waves as the authentication signal. The frequency of the sine wave is switched frequently to keep it from being detected by the attacker. The time period over which the frequency of the sine wave is kept constant is called a *frame*. Watermarking signal over a frame is given by the following equation [44].

$$m(t) = \sum_{i=1}^{n_m} A_i \sin(\omega_i t + \phi_i) \qquad 0 < t < T_f \qquad (2.1)$$

where $t$ is measured from the start of frame, $n_m$ is the number of unique frequencies within a frame and $T_f$ is the frame length. $A_i$, $\omega_i$ and $\phi_i$ represent the amplitude, frequency and phase of the sine

components, respectively. At the end of each frame, another frame with another set of sinusoids is used for watermarking.

Let $f_i = \frac{\omega_i}{2\pi}$ and $T_i = \frac{1}{f_i}$ denote the frequency (in $Hz$) and period (in $s$) of each sinusoidal component in $m(t)$. To simplify the design, the frequencies are chosen such that:

$$\frac{f_1}{n_1} = \frac{f_2}{n_2} = \cdots = \frac{f_{n_m}}{n_{n_m}} \tag{2.2}$$

where $n_1, n_2, \ldots, n_{n_m}$ are relatively prime integers. This ensures that $m(t)$ is a periodic signal with period $T_{comb} = n_1 T_1 = n_2 T_2 = \cdots = n_{n_m} T_{n_m}$. The size of each frame for watermarking is chosen to be a multiple of $T_{comb}$. Therefore, $T_f = kT_{comb}$ for some positive integer $k$.

*Example* 2.1*:* Suppose we have a multi-sine signal defined by three sinusoids at frequencies 10 $Hz$, 20 $Hz$, and 50 $Hz$. The frame will contain these components superimposed and captured as:

$$m(t) = A_1 \sin(2\pi \times 10 \times t + \phi_1) + A_2 \sin(2\pi \times 20 \times t + \phi_2) + A_3 \sin(2\pi \times 50 \times t + \phi_3)$$

$$f_1 = 10\ Hz \text{ or } \omega_1 = 2\pi \times 10 \text{ and } T_1 = \frac{1}{f_1} = \frac{2\pi}{\omega_1} = 0.1\ second \qquad n_1 = 1$$

$$f_2 = 20\ Hz \text{ or } \omega_2 = 2\pi \times 20 \text{ and } T_2 = \frac{1}{f_2} = \frac{2\pi}{\omega_2} = 0.05\ second \qquad n_2 = 2$$

$$f_3 = 50\ Hz \text{ or } \omega_3 = 2\pi \times 50 \text{ and } T_3 = \frac{1}{f_3} = \frac{2\pi}{\omega_3} = 0.02\ second \qquad n_3 = 5$$

For proper detection, the minimum length of frame size should be long enough to accommodate at least one complete length of sine wave of every frequency.

$$T_{comb} = n_1 T_1 = n_2 T_2 = n_3 T_3 = 0.1\ s \tag{2.3}$$

Therefore, $T_{comb} = 0.1\ second$ is the required minimum frame length for the signal considered in this example.

When switching from one frame of sine frequencies to another with a different set of sine frequencies, an abrupt shift from one sinusoidal signal to another forces the output signal to undergo a transient response until the new steady state is reached. From a frequency-domain perspective, this sudden change introduces unintended frequency components which show up as

transient disturbances in the output. [44] shows that by properly choosing the amplitudes, phases and number of unique frequencies in each frame, the transient responses can be suppressed as explained below.

Suppose

$$G_{ym}(s) = \frac{G(s)}{1+K(s)G(s)} \tag{2.4}$$

is an $n^{th}$ order strictly proper system described by the differential equation

$$\frac{d^n y}{dt^n} + a_{n-1}\frac{d^{n-1}y}{dt^{n-1}} + \cdots + a_0 y(t) = b_{n-1}\frac{d^{n-1}m}{dt^{n-1}} + \cdots + b_0 m(t)$$

Therefore,

$$G_{ym}(s) = \frac{b(s)}{a(s)} = \frac{b_{n-1}s^{n-1}+\cdots+b_1 s+b_0}{s^n+a_{n-1}s^{n-1}+\cdots+a_1 s+a_0} \tag{2.5}$$

Also, suppose the input is a multi-sine pulse

$$m(t) = \begin{cases} 0 & t < 0 \text{ and } t > T_f \\ \sum_{i=1}^{n_m} A_i \sin(\omega_i t + \phi_i) & 0 < t < T_f \end{cases} \tag{2.6}$$

To achieve transient suppression, it is shown in [44] that $M(s)$ can be chosen as

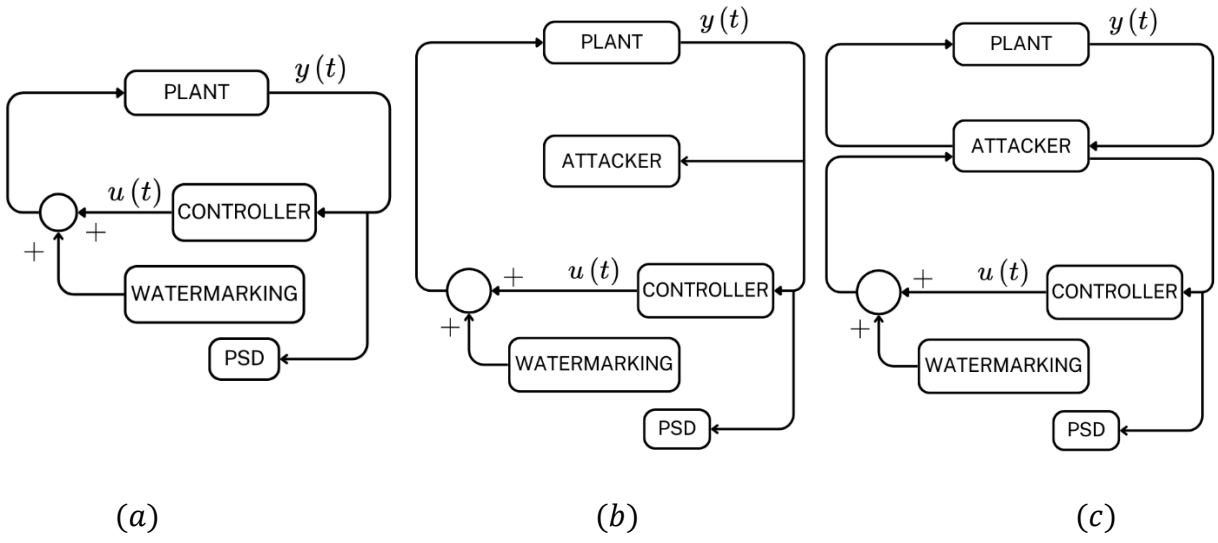
$$M(s) = \frac{c(s)a(s)}{(s^2+\omega_1^2)\ldots(s^2+\omega_{n_m}^2)} \tag{2.7}$$

where $c(s)$ is any polynomial in $s$ and the number of sinusoids $n_m$ is chosen so that

$$n_m \geq n + \frac{deg(c(s))}{2} \tag{2.8}$$

Once $M(s)$ is chosen, the amplitudes $A_i$ and phases $\phi_i$ will be known. A straightforward choice for $c(s)$ is $c(s) = c_0$ and scalar $c_0$ can be used to adjust the power of watermarking signal $m(t)$ so that the watermark frequencies can be detected using periodogram. If $c(s) = c_0$ is used, then (2.7) reduces to $n_m \geq n$.

The periodogram, an estimate of the power spectral density (PSD), of the output signal received by the controller ($Fig.2.4$) is used for analyzing its frequency content and detecting the multi-sine watermarking signal.




$Fig.2.4$. Block diagram of system with watermarking detection using PSD $(a)$ under safe conditions $(b)$ being recorded by attacker and $(c)$ under a replay attack.

The frequency resolution of a periodogram refers to the smallest frequency difference that can be distinguished in the spectral estimate. It determines how well two closely spaced frequency components can be resolved. The frequency resolution is influenced by several factors, primarily the length of the signal and the sampling rate. Let $N$ denote the number of samples and $f_s$ the sampling frequency. The frequency resolution of a periodogram is given by:

$$\Delta f = \frac{f_s}{N} = \frac{1}{NT_s} \tag{2.9}$$

Since the periodogram is computed for one single frame at a time, in multi-sine watermarking a longer frame length (larger $N$) results in finer frequency resolution. The frequency resolution for frame is given by:

$$\Delta f = \frac{1}{T_f} \tag{2.10}$$

where $T_f$ is the size (in $sec$) of the frame. We know from (2.3) that

$$T_f = kT_{comb} = kn_1T_1 = kn_2T_2 = \cdots = kn_{n_m}T_{n_m} \tag{2.11}$$

$$\Delta f = \frac{1}{kT_{comb}} = \frac{f_1}{kn_1} = \frac{f_2}{kn_2} = \cdots = \frac{f_{n_m}}{kn_{n_m}} \tag{2.12}$$

$n_1, n_2, \ldots, n_{n_m}$ are relatively prime integers and suppose they are in ascending order. Therefore, $f_{n_m}$is chosen close to the highest frequency that can be detected within a frame. The numbers $n_1, n_2, \ldots, n_{n_m}$ are kept low such that the frequencies $f_1, f_2, \ldots f_{n_m}$ are appropriately spaced from each other for successful detection. It can be shown that for any $k \geq 1$ [44],

$$f_{i+1} - f_i \geq \Delta f \tag{2.13}$$

We know that the borderline case $f_{i+1} - f_i = \Delta f$ occurs when $n_{i+1} - n_i = 1$. To avoid this, we must take $k \geq 2$. Hence, the frame size $T_f \geq 2T_{comb}$.

*Example* 2.2*:* Consider $n_1 = 3$, $n_2 = 5$, $n_3 = 7$ and $n_4 = 8$ are four relatively prime numbers chosen for the $i^{th}$ frame consisting of four unique frequencies. If $100\ Hz$ is the highest possible frequency that is considered in this frame, we can find the values of other frequencies as follows:

$$f_4 = 100\ Hz$$

$$f_3 = 100\ Hz \times \frac{7}{8} = 87.5\ Hz$$

$$f_2 = 100\ H \times \frac{5}{8} = 62.5\ Hz$$

$$f_1 = 100\ H \times \frac{3}{8} = 37.5\ Hz$$

$T_4 = 0.01s$, $n_4 = 8$, $T_{comb} = 0.08s$ and therefore, $T_f = 0.08k\ s$ is the frame size. Further, we must have

$$\Delta f = \frac{1}{0.08k} \leq f_2 - f_1 = 25\ Hz$$

$$\frac{1}{0.08k} \leq f_3 - f_2 = 25\ Hz$$

$$\frac{1}{0.08k} \leq f_4 - f_3 = 12.5\ Hz$$

Therefore, $k \geq 2$. We can choose any $k$, however, $k$ is preferred to be kept small to have shorter frames. Selecting lower values for the watermarking signal frequencies $f_i$ results in a longer combined period $T_{comb}$ and a larger frame size $T_f$. This extended period increases the time available for an adversary to detect the watermark and adapt their strategies accordingly. Conversely, increasing the frequencies introduces additional challenges. At higher frequencies, the $SNR$ tends to degrade due to a reduction in the system's frequency response magnitude. To mitigate this SNR reduction, higher amplitude $A_i$ watermarking signals may be employed; however, this approach can be undesirable as it increases the watermark's visibility and potential intrusiveness.

Additionally at high frequencies there is greater modeling uncertainty, meaning the accuracy of the system's behavior predictions decreases. This can lead to lower accuracy in the designed watermarking signal. Therefore, there is a trade-off between the choice of lower frequencies and higher frequencies for watermarking.

By estimating the power spectral density (PSD), a periodogram reveals the distribution of signal power across different frequency components, allowing for the detection of periodic structures or hidden patterns within noisy data. A brief review of PSD and some of its estimation methods is provided in the following section.

## 2.4 Power Spectral Density using Periodogram

The PSD estimate of the sampled output is computed to characterize its frequency content and power distribution. PSD is a function that expresses how the power of a signal is distributed across different frequencies. It is often used in signal processing to analyze the frequency content of signals that may not be purely periodic, such as noise or time-varying signals. The PSD shows the power present in a signal as a function of frequency, which makes it useful for identifying the dominant frequencies or patterns within the signal. A discussion of PSD estimation is presented in this section since we used it in analyzing watermarks to detect replay attacks.

For a continuous-time real-valued signal $x(t)$, the average power is

$$P = \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} |x(t) w_T(t)|^2 \, dt \tag{2.14}$$

where

$$w_T(t) = \begin{cases} 1 & -\frac{T}{2} \le t \le \frac{T}{2} \\ 0 & otherwise \end{cases} \tag{2.15}$$

We define

$$x_T(t) = x(t) w_T(t) \tag{2.16}$$

It follows from Parseval's Theorem that

$$P = \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} |\mathcal{F}\{x_T(t)\}|^2 \, df \tag{2.17}$$

where $F\{x_T(t)\}$ is the Fourier Transform of $x_T(t)$ which we will denote as

$$X_T(f) = \mathcal{F}\{x_T(t)\}$$

The PSD of $x(t)$ which shows the density of power in various frequencies is defined as [45]:

$$S_{xx}(f) = \lim_{T\to\infty} \frac{1}{T} |X_T(f)|^2 \tag{2.18}$$

If $x(t)$ is ergodic, then using Weiner-Khinchin theorem we can show that

$$S_{xx}(f) = \int_{-\infty}^{\infty} r_{xx}(\tau) e^{-j2\pi f\tau} \, d\tau = R_{xx}(f) \tag{2.19}$$

where $r_{xx}(\tau)$ is the autocorrelation function.

$$r_{xx}(\tau) = E(x(t+\tau)x(t)) \tag{2.20}$$

and

$$R_{xx}(f) = \mathcal{F}\{r_{xx}(\tau)\}$$

By calculating the PSD, one can determine the power density at each frequency. This is important in many applications like telecommunications, audio processing, and control systems, where understanding the frequency distribution is critical to signal analysis and design.

A periodogram is an estimate of the PSD of a signal. Suppose continuous-time signal $x(t)$ is sampled with sampling period $T_s$ and denote the samples $x[n] = x(nT_s)$ with $0 \leq n \leq N-1$. Using $N$ samples the periodogram is calculated as

$$P_{xx}(f) = \frac{T_s}{N} |X_N(f)|^2 \tag{2.21}$$

where $X(f)$ is the discrete-time Fourier transform of $x[n]$ assuming $x[n] = 0$ for $n < 0$ and $n > N$.

$$X_N(f) = \sum_{n=0}^{N-1} x[n] e^{-j2\pi f n T_s} \tag{2.22}$$

*Remark 2.1*: Consider a signal $y[n]$ obtained from $x[n]$ and padded with $L$ zero values:

$$y[n] = \begin{cases} x[n] & 0 \leq n \leq N-1 \\ 0 & N \leq n \leq N+L-1 \end{cases} \tag{2.23}$$

Then, $y[n]$ has $N + L$ sample values. The periodogram of $y[n]$ is similar to that of $x[n]$ except that it is scaled by $\frac{N}{N+L}$.

$$P_{yy}(f) = \frac{1}{N+L} \sum_{n=0}^{N+L-1} y[n] e^{-j2\pi f n T_s} \tag{2.24}$$

$$P_{yy}(f) = \frac{N}{N+L} \times \frac{1}{N} \sum_{n=0}^{N-1} x[n] e^{-j2\pi f n T_s} \tag{2.25}$$

$$P_{yy}(f) = \frac{N}{N+L} P_{xx}(f) \tag{2.26}$$

*Remark 2.2*: The resolution of a periodogram depends on the length of the sampled signal. Two sinusoids with the same amplitude can be resolved if their frequencies are apart by at least $\frac{1}{NT_s}$. In other words, the resolution is [45]

$$\Delta f = \frac{1}{NT_s} \tag{2.27}$$

When the signal contains noise, the resulting periodogram will be stochastic with variance that does not decrease even when $N \to \infty$. *Welch's method* is a popular technique for estimating the PSD of a signal, serving as an improvement over the traditional periodogram. It reduces the variance of the spectral estimate by averaging multiple periodograms that are computed from overlapping segments of the signal, providing a smoother and more accurate estimate of the PSD. The key concept behind Welch's method is segmenting the signal, windowing each segment, calculating periodograms for each windowed segment, and averaging the results. The main steps in Welch's Method are listed as follows.

1. Segmenting the Signal: The signal is divided into $M$ overlapping segments. If the total length of the signal is $N$, and the segment length is $L$, each segment typically overlaps by 50% or more. Overlapping ensures that more data points contribute to each frequency estimate, reducing the variability of the estimate across different realizations of the signal. This leads to a more stable and consistent PSD estimate by reducing fluctuations caused by the randomness of individual segments.

2. Windowing: Each segment is multiplied by a window function to reduce spectral leakage. Windowing tapers the ends of each segment to zero, minimizing the discontinuities at the boundaries of the segments.

3. Computing the Periodogram: For each windowed segment, the periodogram is computed using the DFT or FFT. The periodogram for each segment is given by:

$$P_{xx}^{k}(f) = \frac{T_s}{L}\left|\sum_{n=0}^{L-1} x_m[n]w[n]e^{-\frac{j2\pi f n}{M}T_s}\right|^2 \tag{2.28}$$

where $x_m[n]$ is the $m^{th}$ segment of the signal, $w[n]$ is the window function, $M$ is the length of each segment and $f$ represents the frequency.

4. Averaging: The periodograms from all the segments are averaged to produce a final PSD estimate. This averaging reduces the variance of the PSD, yielding a more reliable estimate. The Welch's estimate of the PSD is given by:

$$\widehat{P_{xx}}(f) = \frac{1}{M}\sum_{m=1}^{M} P_{xx}^{k}(f) \tag{2.29}$$

where $\widehat{P_{xx}}(f)$ is the Welch PSD estimate, $M$ is the number of overlapping segments and $P_{xx}^{m}(f)$ is the periodogram of the $m^{th}$ segment.

Averaging multiple periodograms reduces the variance of the PSD estimate compared to a single periodogram, making it more reliable for noisy or nonstationary signals. By adjusting the segment length, overlap, and window function, Welch's method offers flexibility in balancing frequency resolution and smoothing. Shorter segments provide better smoothing but reduce frequency resolution, while longer segments improve resolution but may increase variance.

## 2.5 Confidence Interval of Periodogram

As previously mentioned, if $x[n]$ is a random process, the periodogram will be a random process. It can be shown that the expected value of $P_{xx}(f)$ tends to the PSD when $N \to \infty$ [45]-[46] is

$$\lim_{N\to\infty} E\big(P_{xx}(f)\big) = S_{xx}(f) \tag{2.30}$$

Further, the variance of $P_{xx}(f)$ (when $f$ is not close to 0 or $\pm\frac{1}{2T_s}$) is

$$var(P_{xx}(f)) \approx S_{xx}(f) \tag{2.31}$$

Note that the variance does not depend on $N$ and could be high even when $N \to \infty$. One way to reduce the variance is to divide the $N$ samples of $x[n]$ into $M$ segments of length $L$ ($N = ML$), find the periodogram of each segment $P_{xx}^{m}(f)$ with $m = 0, \dots, M-1$ and then form the average periodogram.

$$P_{xx,avg}(f) = \frac{1}{M}\sum_{m=0}^{M-1} P_{xx}^{m}(f) \tag{2.32}$$

If $x[n]$ is WSS Gaussian, then data blocks will be approximately uncorrelated and the periodograms will be independent. Thus

$$var\left(P_{xx,avg}(f)\right) = \frac{1}{M} var(P_{xx}^{m}(f)) \tag{2.33}$$

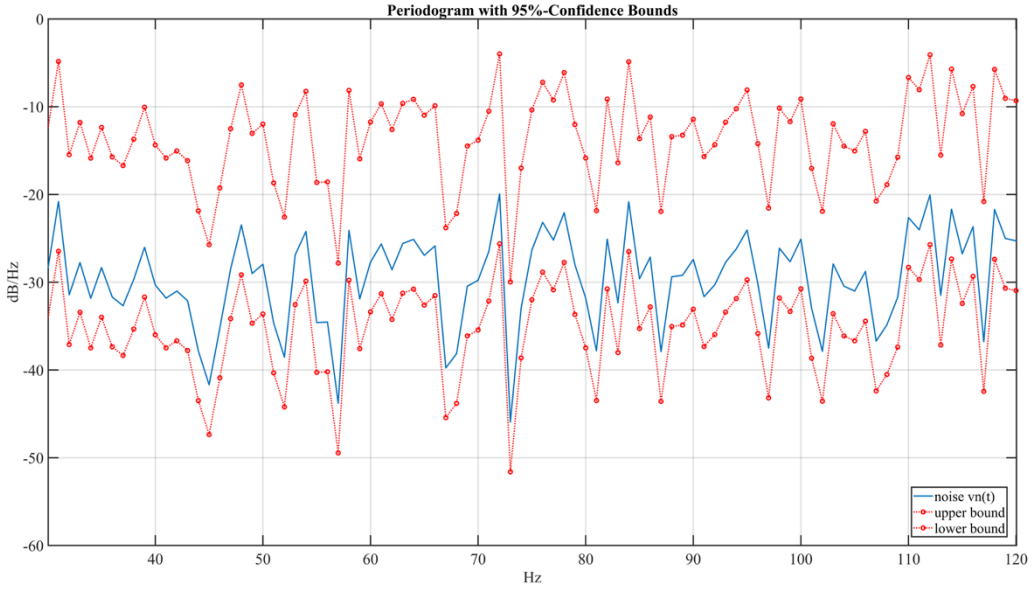
If $x[n]$ is WSS Gaussian, then $X_N(f)$ will be WSS Gaussian and that $\frac{2MP_{xx,avg}(f)}{S_{xx}(f)}$ will have a $\chi^2_{2M}$ distribution [45]. Thus, the $\alpha\%$ interval of confidence will be

$$\frac{2MP_{xx,avg}(f)}{\chi^2_{2M}\left(\frac{\alpha}{2}\right)} < S_{xx}(f) < \frac{2MP_{xx,avg}(f)}{\chi^2_{2M}\left(1-\frac{\alpha}{2}\right)} \tag{2.34}$$

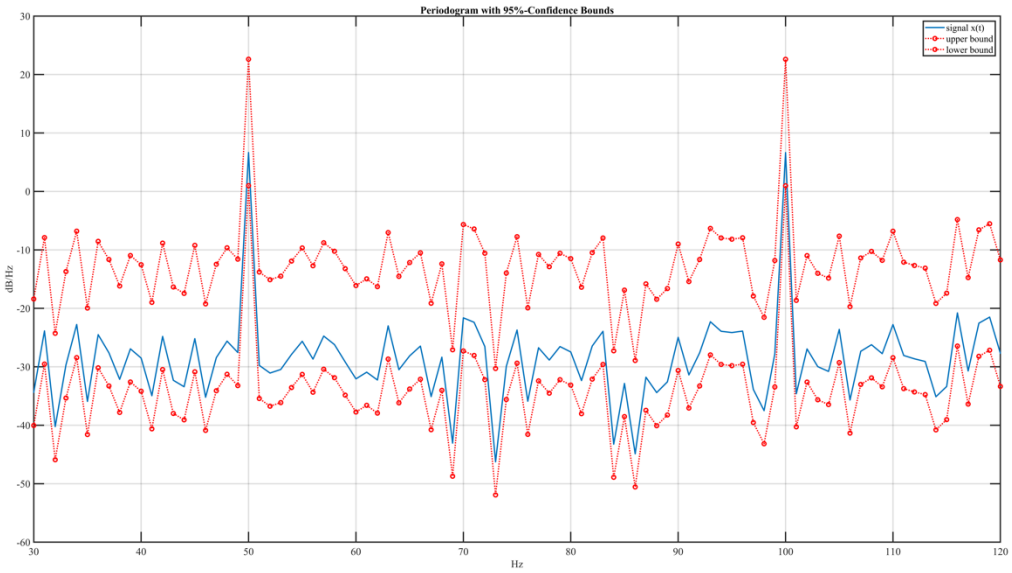
*Example 2.2*: Consider a process:

$$x(t) = 3\sin(2\pi f_1 t + \theta_1) + 3\sin(2\pi f_2 t + \theta_2) + v_n(t)$$

where $f_1 = 50\ Hz$, $\theta_1 = \frac{\pi}{6}$, $f_2 = 100\ Hz$, $\theta_2 = 0$ and $v_n$ is Gaussian noise with zero mean and variance $\sigma_v^2 = 1$. Suppose $f_s = 1000\ Hz$. The periodogram with 95% confidence bounds of noise $v_n(t)$ and $x(t)$ are shown in $Fig\ 2.5(a)$ and $Fig.\ 2.5(b)$, respectively.




$(a)$




$(b)$

*Fig.* 2.5. *Example 2.2:* 95% confidence bound periodogram of $(a)\ v_n(t)\ (b)\ x(t)$.

We can see in *Fig.* 2.5$(b)$ that the lower bound estimates at $f = 50\ Hz$ and $f = 100\ Hz$ are higher than the upper bound estimates at the other frequencies which points to the two sinusoidal components with a confidence of 95%. This demonstrates the utility of confidence bounds in the context of using a periodogram for detection of frequencies.

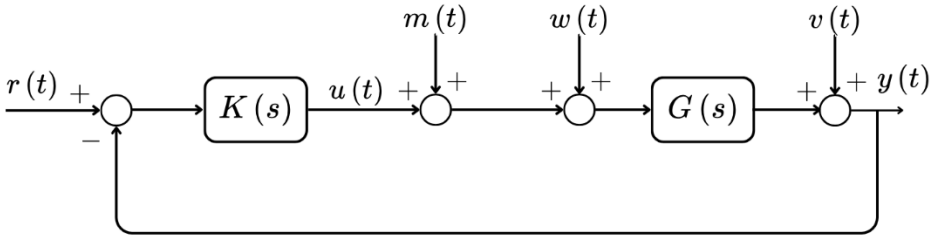
# Chapter 3

# Multi-Sine Watermark Design

In this chapter, we develop a methodology for choosing the parameters of the multi-sine watermarking signal. The detection of attack is based on the power spectral density of the received output signal. Thus, in this chapter, we first provide an analysis of the PSD before and during a replay attack. Next, using our analysis, we present our algorithm for choosing watermark parameters to achieve the desired confidence level in attack detection.

## 3.1 Problem Statement

Consider the LTI system with watermarking shown in $Fig.3.1$. Watermarking starts when the closed loop system reaches steady state.




$Fig$ 3.1. Block diagram of LTI system.

In $Fig.3.1$, as described in chapter 2, $K(s)$ is the controller transfer function, $G(s)$ the plant transfer function, $r(t)$ the reference input signal, $u(t)$ the control signal, $m(t)$ the watermarking signal, $w(t)$ the input disturbance, $v(t)$ the measurement noise and $y(t)$ the measured output signal. Signals $w(t)$ and $v(t)$ are considered to be IID Gaussian with mean and variance given by $(0, \sigma_w^2)$ and $(0, \sigma_v^2)$, respectively. The closed loop system is assumed to be stable and in steady state. We consider multi-sine watermarking. Let the multi-sine watermarking signal over a frame be given as:

$$m(t) = \sum_{i=1}^{n_m} A_i \sin(\omega_i t + \phi_i) \tag{3.1}$$

where $t$ is measured from the start of frame, $n_m$ is the number of unique frequencies within a frame and $T_f$ is the frame length. Furthermore, $A_i$, $\omega_i$ and $\phi_i$ represent the amplitude, frequency and phase of the sine components, respectively. The frequencies in $Hz$ and periods are denoted by

$$f_i = \frac{\omega_i}{2\pi} \tag{3.2}$$

$$T_i = \frac{1}{f_i} \tag{3.3}$$

As discussed in section 2.3, within a frame, the frequencies are chosen such that

$$\frac{f_1}{n_1} = \frac{f_2}{n_2} = \cdots = \frac{f_{n_m}}{n_{n_m}} \tag{3.4}$$

for positive integers $n_1, \dots, n_{n_m}$ with $GCD(n_1, n_2, \dots n_{n_m}) = 1$. Thus, with $T_{comb}$ defined as

$$T_{comb} = n_1 T_1 = n_2 T_2 = \cdots = n_{n_m} T_{n_m}, \tag{3.5}$$

the frame length is chosen as

$$T_f = kT_{comb} \tag{3.6}$$

for some $k \geq 2$. The watermark signal is chosen to be

$$M(s) = \frac{c(s)a(s)}{(s^2+\omega_1^2)\dots(s^2+\omega_{n_m}^2)} \tag{3.7}$$

with

$$n_m \geq n + \frac{1}{2}\deg\left(c(s)\right) \tag{3.8}$$

to do watermarking and prevent generating transient response. Assuming

$$c(s) = c_o \qquad \text{(i.e. } \deg\big(c(s)\big) = 0\text{)} \tag{3.9}$$

(3.8) becomes

$$n_m \geq n. \tag{3.10}$$

When the multi-sine watermarking signal is added to the plant input, it introduces a known, low-power sinusoidal signal into the output response. This watermarking signal is designed to be

distinguishable from noise, allowing it to be detected and analysed without significantly affecting the system's performance. Its presence helps in monitoring, identification, and validation of the system's behaviour. Therefore, the watermarking signal must be chosen to achieve the following main goals:

$G1$) In the absence of replay attack, the watermark's presence in the system output must be detectable with sufficient confidence despite the effects of plant disturbance and measurement noise.

$G2$) During a replay attack, the watermark's change or absence in the system output must be detectable with sufficient confidence.

To address the above, we examine the PSD estimations and the impact of the choices of the following parameters of $m(t)$:

$n_f$: number of frames

$n_m$: number of watermark sinusoids in each frame

$f_i$: watermark frequencies for each frame

$c_o$: scaling factor to adjust the power of $m(t)$

[44] provides certain conditions to design some of the parameters, namely, (3.4), (3.6) and (3.8). In this chapter, we examine the PSD before and during replay attack to obtain an algorithm for choosing the above watermarking parameters. Our objective is to satisfy goals $G1$ and $G2$. We start with an analysis of PSD of output before replay attack.

## 3.2 Watermark Detection Before Replay Attack

Choosing the right parameters is important to ensure that the watermark is detected in the output, remains hidden from the attacker, and does not exceed the output fluctuation limits. The selection process balances a trade-off between strength of the watermarking signal and output signal regulation quality, i.e. to minimize output fluctuations while keeping the watermark detection reliable. The following discussion outlines the criteria and methods used to determine the optimal parameter values for effective watermarking before replay attack.

The multi-sine watermarking signal results in fluctuations in plant output. The amplitudes $A_i$ should be chosen so that the output perturbations are small enough that do not degrade the quality of output regulation but at the same time, the output perturbations should be large enough for watermark to be detected and distinguished from noise. The periodogram provides insight into the frequency content of the output signal, helping us to identify dominant frequencies and noise characteristics. Confidence intervals help quantify the uncertainty associated with the periodogram estimates, providing a range within which the true PSD is likely to lie. This is important because it allows us to assess the reliability of the frequency estimates and make more informed decisions when designing or tuning the watermark. A narrower confidence interval indicates higher precision in the spectral estimate while a wider interval suggests greater variability or noise in the data. By considering confidence intervals, we can better distinguish between meaningful signal components and random noise, improving the robustness and accuracy of watermark detection. Confidence interval for periodograms are available, as discussed in Section 2.5, for wide-sense stationary (WSS) IID Gaussian signals. We show that in the system under our consideration ($Fig.\,3.1$), the noise present in the output is WSS Gaussian.

*Theorem 3.1*. Suppose in the system of $Fig.\,3.1$, $v(t)$ and $w(t)$ are WSS Gaussian with zero mean. Then the noise in the sampled output $y[n]$ will be WSS Gaussian.

*Proof.* To examine the noise in the output, let all inputs other than $v(t)$ and $w(t)$ are zero i.e., $r(t) = 0$ and $m(t) = 0$. Suppose the closed-loop system has reached steady state with transfer functions:

$$G_{yw}(s) = \frac{Y(s)}{W(s)} \tag{3.11}$$

and

$$G_{yv}(s) = \frac{Y(s)}{V(s)} \tag{3.12}$$

Let $g_{yw}(t) = L^{-1}\{G_{yw}(s)\}$ and $g_{yv}(t) = L^{-1}\{G_{yv}(s)\}$. Then the output will be

$$y = y_w + y_v \tag{3.13}$$

$$= g_{yw} * w + g_{yv} * v \tag{3.14}$$

where $*$ denotes convolution. Further, we observe that

$$E\big(y_w(t_1)y_v(t_2)\big) = E\big(\big(g_{yw} * w\big)(t_1)\big(g_{yv} * v\big)(t_2)\big) \tag{3.15}$$

$$= E\Big(\int_{-\infty}^{+\infty} g_{yw}(t)w(t_1 - t)dt \int_{-\infty}^{+\infty} g_{yv}(t')v(t_2 - t')dt'\Big) \tag{3.16}$$

$$= E\Big(\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g_{yw}(t)g_{yv}(t')w(t_1 - t)v(t_2 - t')dt\,dt'\Big) \tag{3.17}$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g_{yw}(t)g_{yv}(t')E\big(w(t_1 - t)v(t_2 - t')\big)dt\,dt' \tag{3.18}$$

$$= 0 \quad \text{(since } w \text{ and } v \text{ are independent with zero mean)} \tag{3.19}$$

Furthermore

$$E\big(y_w(t)\big) = E\big(g_{yw} * w\big) \tag{3.20}$$

$$= E\Big(\int g_{yw}(t)w(t_1 - t)dt\Big) \tag{3.21}$$

$$= 0 \qquad \text{since } E\big(w(t)\big) = 0 \tag{3.22}$$

Similarly,

$$E\big(y_v(t)\big) = 0 \tag{3.23}$$

Signals $y_w(t)$ and $y_v(t)$ are the output of LTI systems with WSS Gaussian inputs. Therefore, they are also WSS Gaussian. Furthermore, (3.22) and (3.23) show that they have zero mean and (3.19) shows that they are uncorrelated and thus independent. As a result, $y_w[n] = y_w(nT)$ and $y_v[n] = y_v(nT)$ are WSS independent Gaussian signals. This implies that $y[n] = y_w[n] + y_v[n]$ is WSS Gaussian.

■

*Example 3.1:* Consider the system given in $Fig.3.1$ with $K(s) = 3$ and $G(s) = \frac{1}{s+1}$. For this system, suppose the watermark is the sinusoidal signal with the frequency $\omega_1 = 1\,rad/s$:

$$M(s) = c_o \frac{s+4}{s^2+1}$$

We know that

$$G_{ym} = \frac{Y(s)}{M(s)} = \frac{G(s)}{1 + K(s)G(s)} = \frac{1}{s+4}$$

Signals $w(t)$ and $v(t)$ are WSS Gaussian with zero mean and variance $\sigma_w^2 = \sigma_v^2 = 1$. We take $\omega_1 = 1\, rad/s$, $T_1 = \frac{2\pi}{\omega_1}$ and $T_f = kT_1 = 4T_1$ (suppose $k = 4$). Also,

$$G_{yw} = \frac{Y(s)}{M(s)} = \frac{1}{s+4}$$

$$G_{yv}(s) = \frac{1}{1 + K(s)G(s)} = \frac{s+1}{s+4}$$

The frame size and sampling period are chosen as $T_f = 8\pi\, s$ and $T_s = 0.01\, s$, respectively. Furthermore, we also consider two different values of $c_o$ to observe how it influences the detection using 95% confidence bound periodogram. First, we choose $c_0 = 0.05$. $Fig.3.2$ shows the output periodogram with 95% confidence intervals with noise set to zero ($w(t) = 0\, and\, v(t) = 0$). $Fig.3.3$ shows the same for $c_0 = 50$. We see that in both $Fig.3.2$ and $Fig.3.3$, the lower bound at $\omega_1 = 1\, rad/s$ is $\sim 13dB$ higher than upper bounds at most other frequencies. Thus, the watermark can be detected with 95% confidence for both values of $c_0$ since there is no noise present.

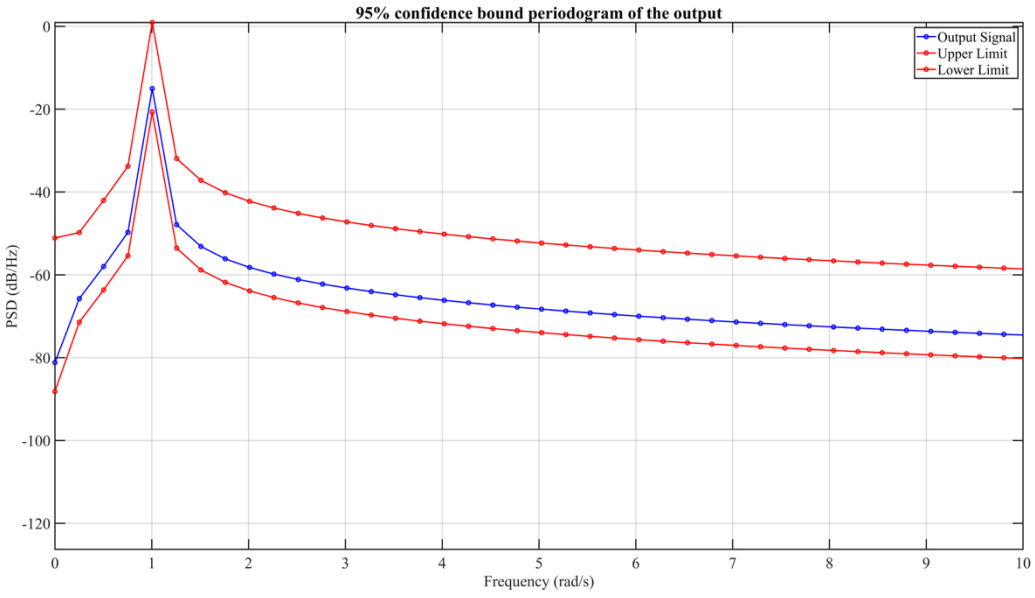



*Fig.3.2. Example 3.1:* Periodogram of output without noise for $c_o = 0.05$.

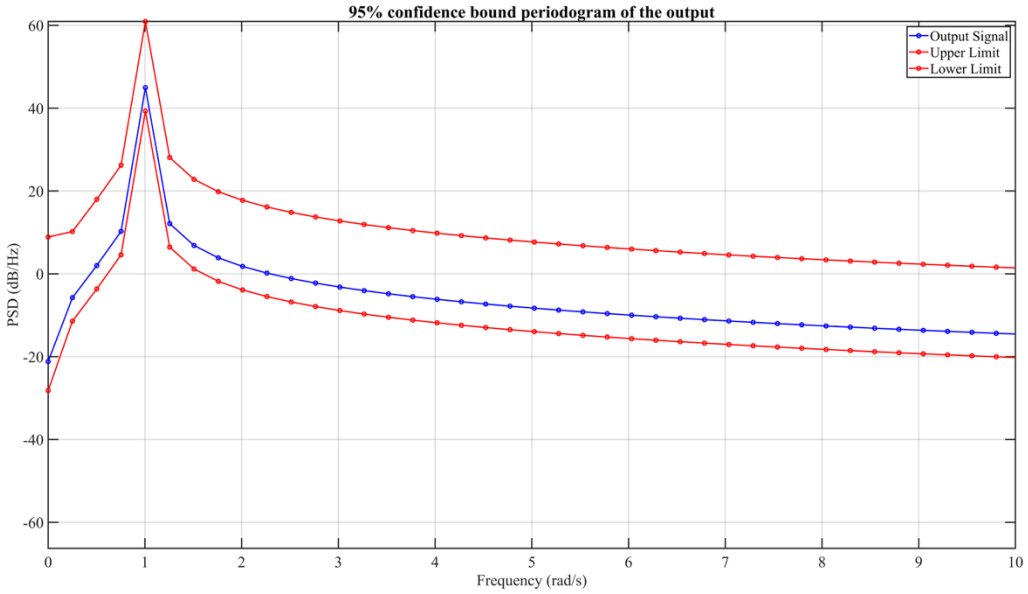



*Fig*. 3.3. *Example 3.1:* Periodogram of output without noise for $c_o = 50$.

Next, we observe the 95% confidence bound periodograms with noise present as shown in *Fig*. 3.4 for $c_0 = 0.05$ and *Fig*. 3.5 for $c_0 = 50$. In *Fig*. 3.4, the frequency is undetectable but in *Fig*. 3.5, for a higher value of $c_0$, the lower bound for $\omega_1 = 1\ rad/s$ goes to ~13 $dB$ higher than the upper bound at other frequencies. Thus, when there is noise present, the sine wave of $\omega_1 = 1\ rad/s$ can be detected with 95% confidence only when $c_0$ is high enough ($c_0 = 50$ in this case).

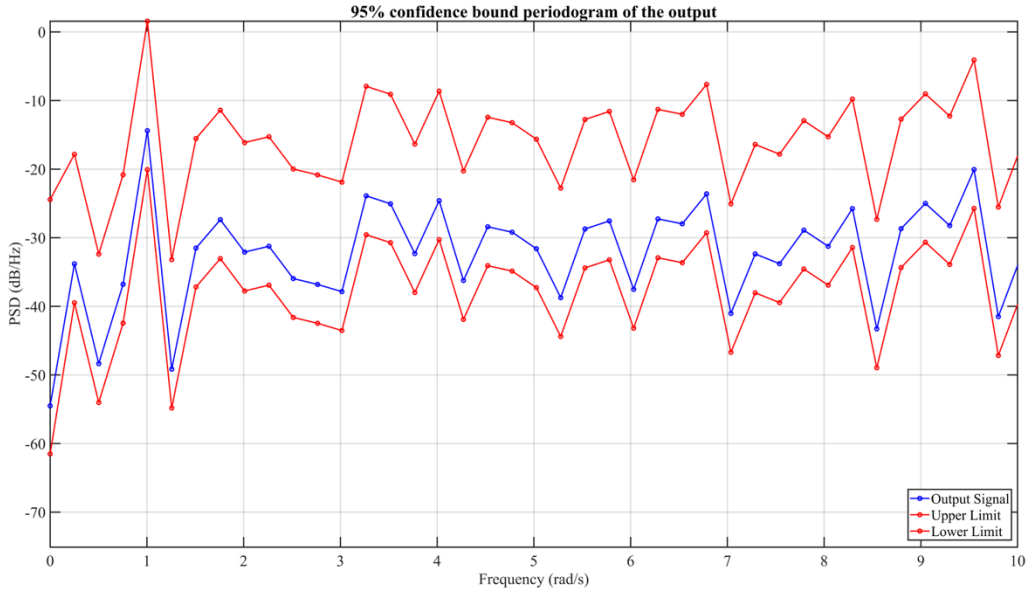



*Fig*. 3.4. *Example 3.1:* Periodogram of output with noise for $c_o = 0.05$.

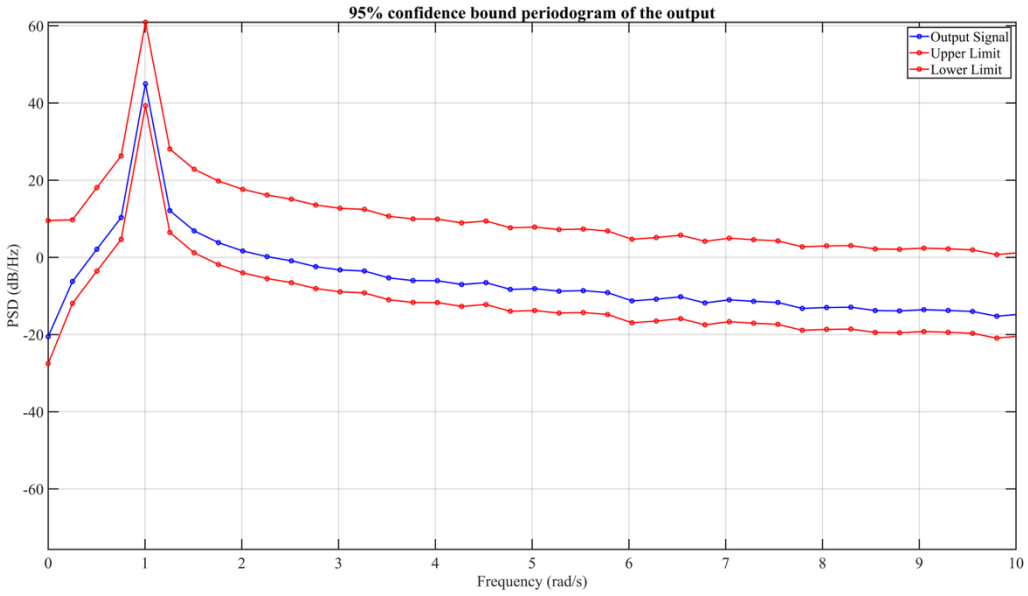



*Fig.* 3.5. *Example 3.1:* Periodogram of output with noise for $c_o = 50$.

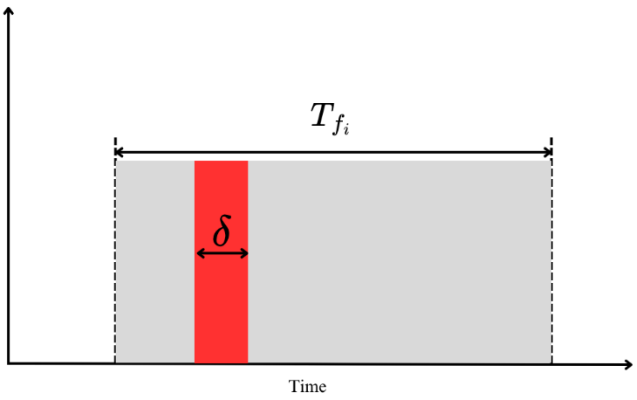
## 3.3 Watermark Changes During Replay Attack

In this section, we discuss the characteristics of the PSD of watermarked output during a replay attack. During attack, a previously recorded portion of output which contains watermarking is replayed repeatedly. Therefore, we assume a replay attack is taking place and the received output signal for the frame under consideration is a replayed signal from other frames.

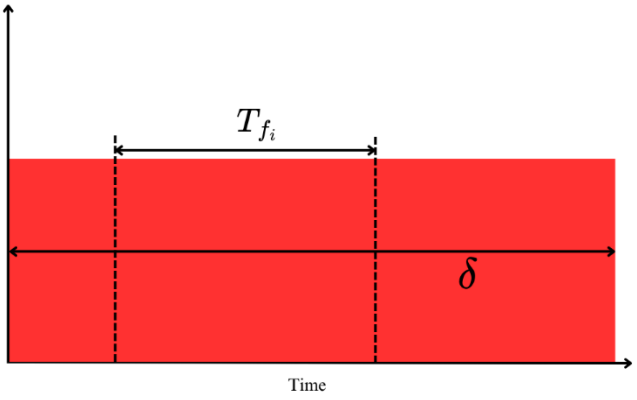
An attacker records and replays a segment of the output. Let $\delta$ denote the length of the recorded segment. Suppose this segment is replayed during a frame $i$ with length $T_{f_i}$ $(i = 1,2, \dots, n_f)$. We consider the following two cases (*Fig.* 3.6):

(1) $\delta \ll T_{f_i}$

(2) $\delta \gg T_{f_i}$




$(a)$




$(b)$

$Fig.$ 3.6. Demonstration of $(a)$ case $(1)$: $\delta \ll T_{f_i}$ and $(b)$ case $(2)$: $\delta \gg T_{f_i}$.

In case $(1)$, the replayed segment is shorter than all frames. Assuming that the frame sizes are of the order of closed loop settling times or longer, in case $(1)$, the repeated segment is shorter than the settling time. This is the case in process industries where settling times are of the order of several minutes and the replayed segment could be of the order of seconds (For instance, in the case of Stuxnet the replayed segment was $21s$ long).

In case $(2)$, the replayed segment is longer than all frames. This case typically occurs in power system applications where settling times are of the order of a second.

### 3.3.1 Short Repeated Segments

Now, we examine the case of $\delta \ll T_{f_i}$. Let $y(t)$ denote the output received by controller and suppose it is a sinusoid with period $T = \frac{2\pi}{\omega_0}$:

$$y(t) = A\sin(\omega_0 t + \phi) \tag{3.24}$$

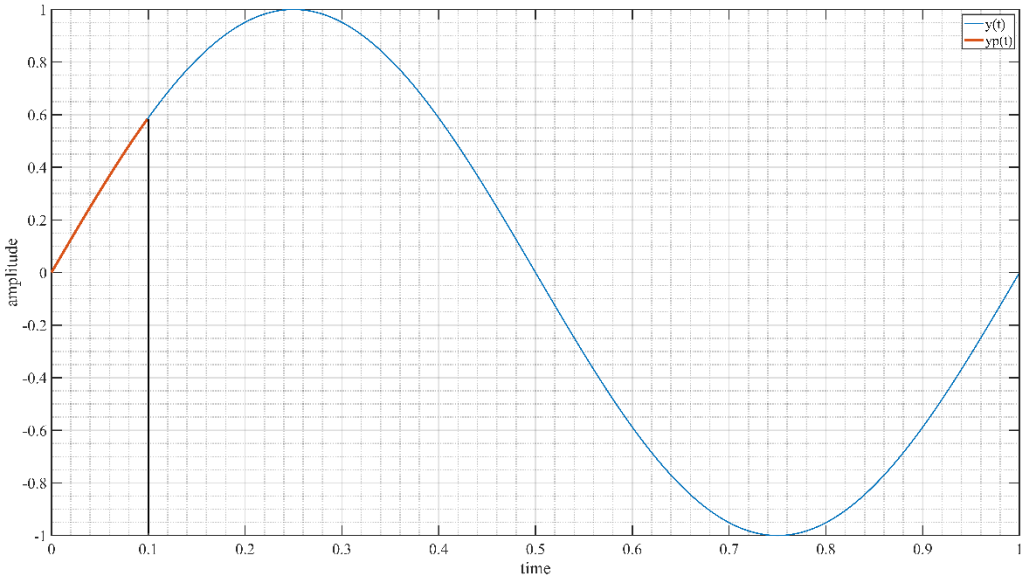
The frame length $T_f$ is a multiple of $T$ and since $\delta \ll T_f$,

$$\delta < T$$

Let the slice of $y(t)$ between $t = 0$ and $t = \delta$ be denoted by $\tilde{y}(t)$ and the signal obtained by replaying $\tilde{y}(t)$ be $\tilde{y}_p(t)$:

$$\tilde{y}_p(t) = y(t - k\delta) \quad 0 < k\delta \leq t < (k+1)\delta \tag{3.25}$$

For example, $Fig.3.7$ shows $y(t)$ with $T_s = 0.001s$, $T_f = T = 1s$ and $\delta = 0.1s$. $Fig.3.8$ shows $\tilde{y}_p(t)$ for $0 \leq t \leq 1s$.




$Fig.3.7$. $\tilde{y}(t)$: a slice of $y(t)$ between $t = 0$ and $t = \delta$ shown in orange.

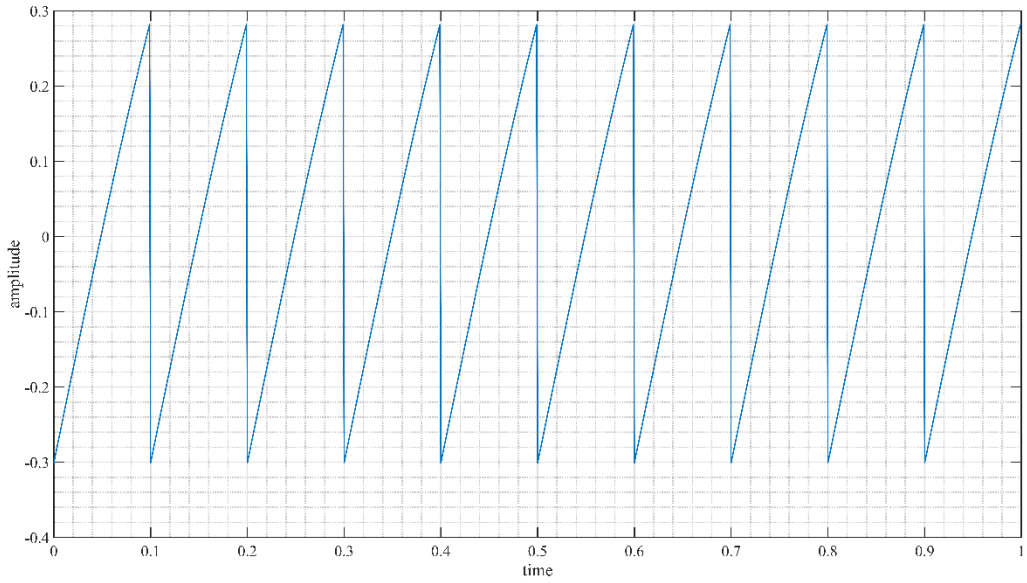



*Fig.* 3.8. $\tilde{y}_p(t)$ obtained by replaying $\tilde{y}(t)$.

$\tilde{y}_p(t)$ is periodic for $t > 0$ and its spectrum contains frequency $\omega_\delta = \frac{2\pi}{\delta}$ and its harmonics $k\omega_\delta$ $(k \geq 2)$. Let us find the power of the main harmonic $\omega_\delta$. Consider the Fourier series of $\tilde{y}_p(t)$:

$$\tilde{y}_p(t) = \sum_{n=-\infty}^{+\infty} c_n e^{jn\omega_\delta t}$$

where

$$c_n = \frac{1}{\delta}\int_0^\delta y(t)e^{-jn\omega_\delta t}dt$$

For $n = 1$,

$$c_1 = \frac{1}{\delta}\int_0^\delta A\sin(\omega_0 t + \phi)\,(\cos\omega_\delta t - j\sin\omega_\delta t)\,dt$$

$$= \frac{A}{\delta}\int_0^\delta (\sin(\omega_0 t + \phi)\cos\omega_\delta t - j\sin(\omega_0 t + \phi)\sin\omega_\delta t)\,dt$$

Further, we observe that

$$\int_0^{\delta} \sin(\omega_0 t + \phi) \cos \omega_\delta t \, dt$$

$$= \int_0^{\delta} \frac{1}{2} (\sin[(\omega_0 + \omega_\delta)t + \phi] + \sin[(\omega_0 - \omega_\delta)t + \phi]) \, dt$$

$$= \frac{1}{2}\left[\frac{-\cos[(\omega_0 + \omega_\delta)t + \phi]\,|_0^{\delta}}{\omega_0 + \omega_\delta} + \frac{\left[-\cos[(\omega_0 - \omega_\delta)t + \phi]\,|_0^{\delta}\right]}{\omega_0 - \omega_\delta}\right]$$

$$= \frac{\omega_0}{\omega_\delta^2 - \omega_0^2} (\cos\phi - \cos(\omega_0\delta + \phi))$$

Similarly

$$\int_0^{\delta} -\sin(\omega_0 t + \phi) \sin \omega_\delta t \, dt$$

$$= \int_0^{\delta} \frac{1}{2} (\cos[(\omega_0 + \omega_\delta)t + \phi] - \cos[(\omega_0 - \omega_\delta)t + \phi]) \, dt$$

$$= \frac{\omega_\delta}{\omega_\delta^2 - \omega_0^2} (\sin\phi - \sin(\omega_0\delta + \phi))$$

Therefore,

$$c_1 = \frac{A}{\delta\left(\omega_\delta^2 - \omega_0^2\right)} (\omega_0(\cos\phi - \cos(\omega_0\delta + \phi)) + j\omega_\delta(\sin\phi - \sin(\omega_0\delta + \phi))$$

If $\delta \ll T$, then $\omega_\delta \gg \omega_0$ and $\omega_\delta^2 - \omega_0^2 \approx \omega_\delta^2$. Thus

$$c_1 \approx \frac{A}{2\pi}\left[\frac{\delta}{T}(\cos\phi - \cos(\omega_0\delta + \phi)) + j(\sin\phi - \sin(\omega_0\delta + \phi))\right]$$

The amplitude of the base harmonic is

$$|c_1| + |c_{-1}| = 2|c_1| = \frac{A}{\pi}\left[\frac{\delta^2}{T^2}(\cos\phi - \cos(\omega_0\delta + \phi))^2 + (\sin\phi - \sin(\omega_0\delta + \phi))^2\right]^{\frac{1}{2}}$$

If $\frac{\delta}{T} \leq 10$, then $\frac{\delta^2}{T^2} \leq 0.01$ and $\sin\left(\frac{\omega_0\delta}{2}\right) \leq \frac{1}{3}$; therefore,

$$|\sin\phi - \sin(\omega_0\delta + \phi)| = 2\left|\cos\left(\phi + \frac{1}{2}\omega_0\delta\right)\right|\left|\sin\frac{\omega_0\delta}{2}\right| \leq \frac{2}{3}$$

$$|\cos\phi - \cos(\omega_0\delta + \phi)| = 2\left|\sin\left(\phi + \frac{1}{2}\omega_0\delta\right)\right|\left|\sin\frac{\omega_0\delta}{2}\right| \leq \frac{2}{3}$$

Therefore, an upper bound for the amplitude of the base harmonic will be

$$\frac{2.2}{3\pi}A \tag{3.26}$$

Thus, the power of the first harmonic drops by $\left(\frac{3\pi}{2.2}\right)^2 = 18$ or $13\ dB$. Note that the original signal and replayed signal have different frequencies. Thus, the small change in periodogram is due to the difference in frequency. $Fig.\,3.9$ shows the periodogram of $y(t)$ and $Fig.\,3.10$ shows that of $\tilde{y}_p(t)$ from $Fig.\,3.8$.

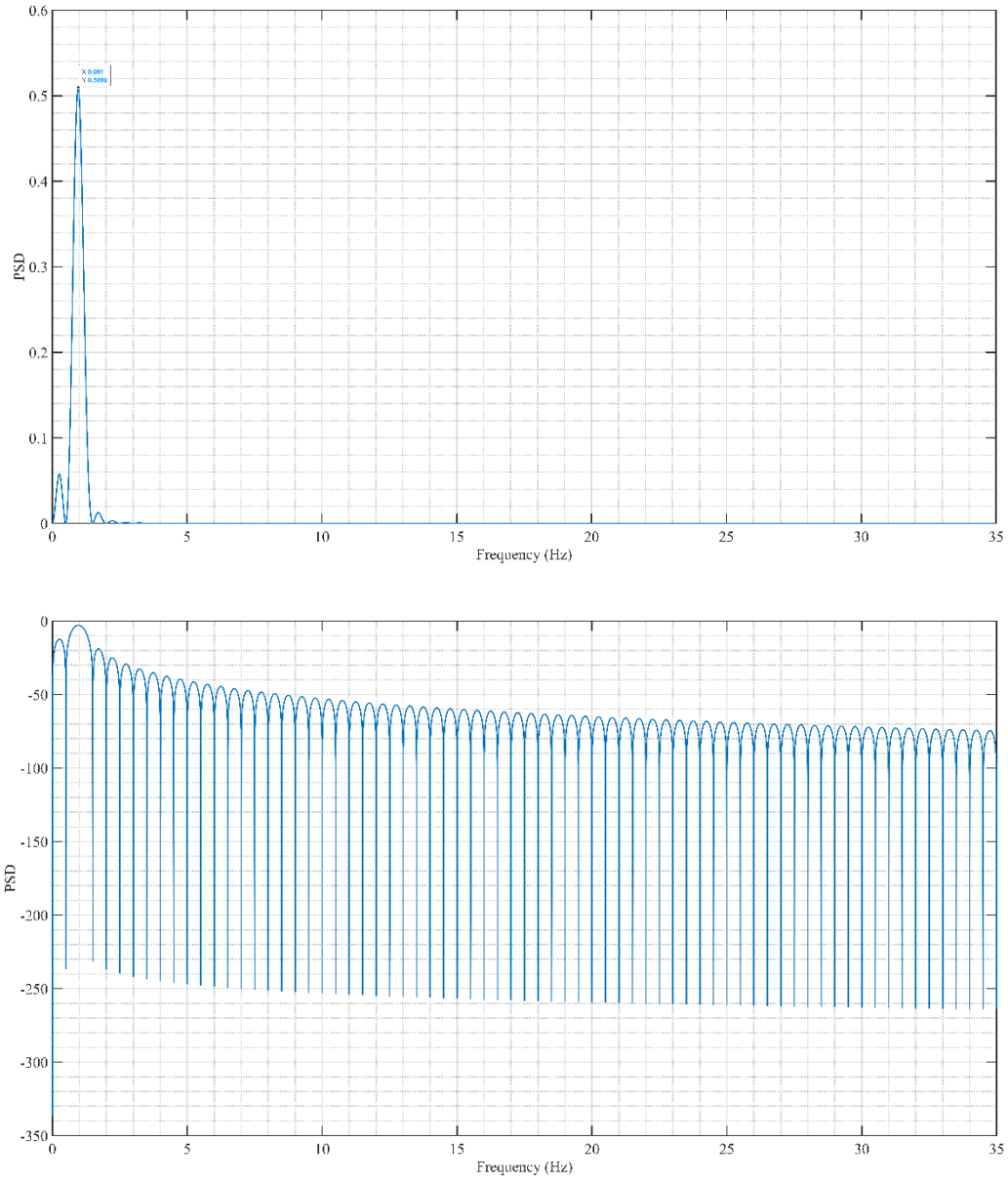
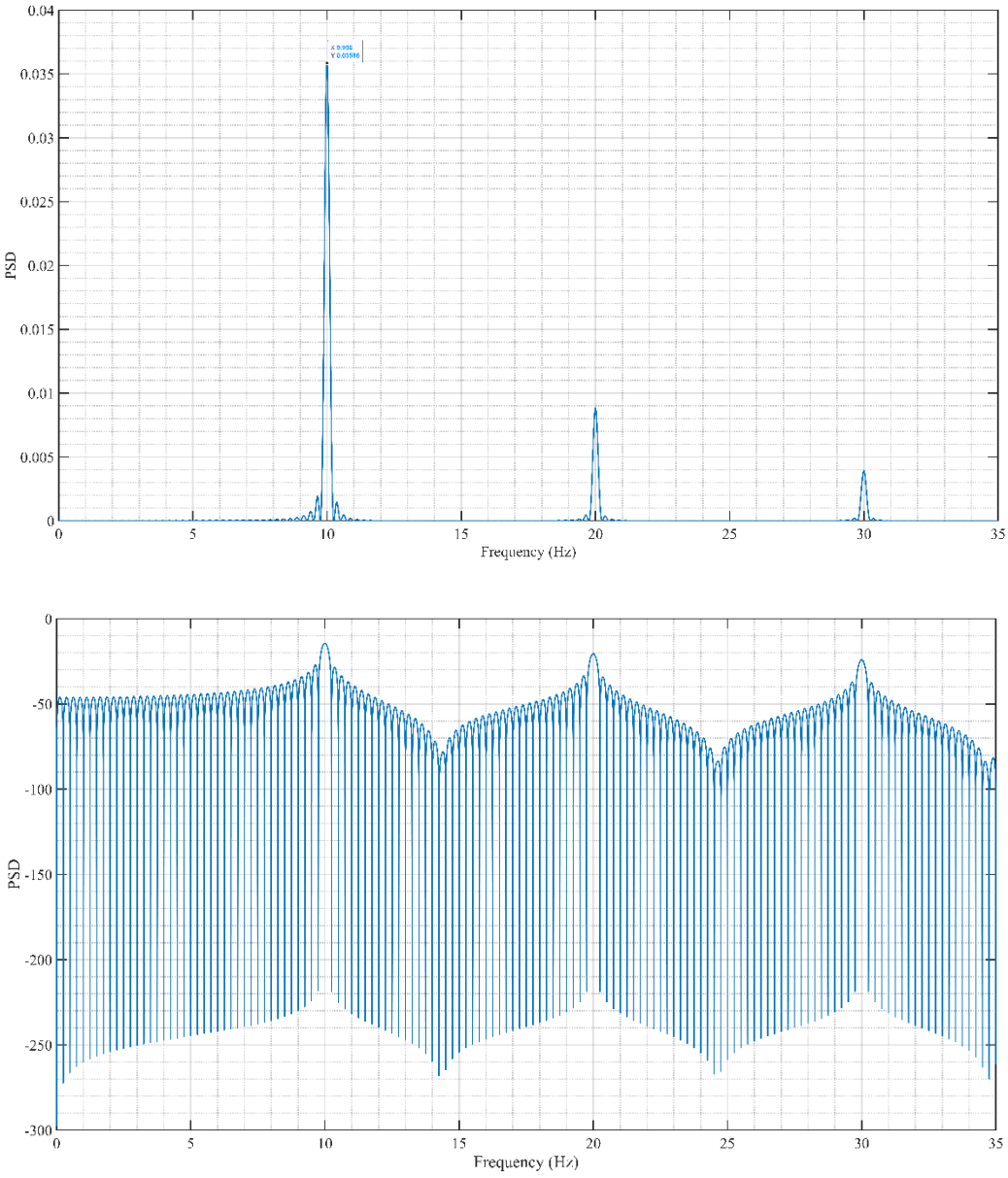



*Fig*. 3.9. Periodogram of $y(t)$ shown in *Fig*. 3.7.




$Fig.$ 3.10. Periodogram of $\tilde{y}_p(t)$ shown in $Fig.$ 3.8.

It can be observed from the above figures that there was a drop from $-2.937\ dB$ to $-14.45\ dB$ in the main harmonic, i.e. $\sim 11.5\ dB$ in the periodogram.

In summary, for case (1) , the periodogram of the frame shows a periodic signal at frequency very much higher than the watermarking signal. Due to the reduction in amplitude that was just discussed, it is possible that the periodic signal becomes indistinguishable from noise.

### 3.3.2 Long Replayed Segments

Now, we turn our attention to the case of $\delta \gg T_{f_i}$. We consider the case where recorded data from some frame $j$ is replayed over another frame $i$. The range of frame sizes are expected to be close and that is why from now on, we assume

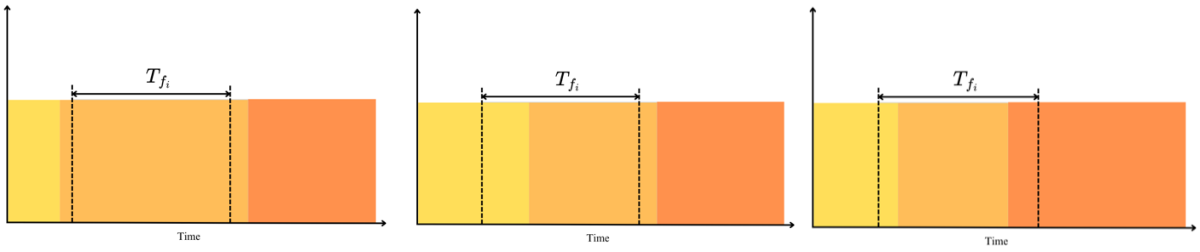
$$\frac{T_{f_i}}{T_{f_j}} < 2 \qquad \forall i, j = 1,2, \dots n_f. \qquad (3.27)$$

This case can be further classified into the following three cases based on the nature of the replayed data as shown in $Fig.\,3.11$.

Case $(2A)$ The replay attack carried over the $i^{th}$ frame contains the replayed signal of a single frame $j$.

Case $(2B)$ The replay attack carried over the $i^{th}$ frame contains the replayed signal of two frames $j$ and $j+1$.

Case $(2C)$ The replay attack carried over the $i^{th}$ frame contains the replayed signal of three frames $j, j+1$ and $j+2$.




(a) (b) (c)

$Fig.\,3.11$. Depiction of $(a)$ Case $(2A)$ $(b)$ Case $(2B)$ $(c)$ Case $(2C)$.

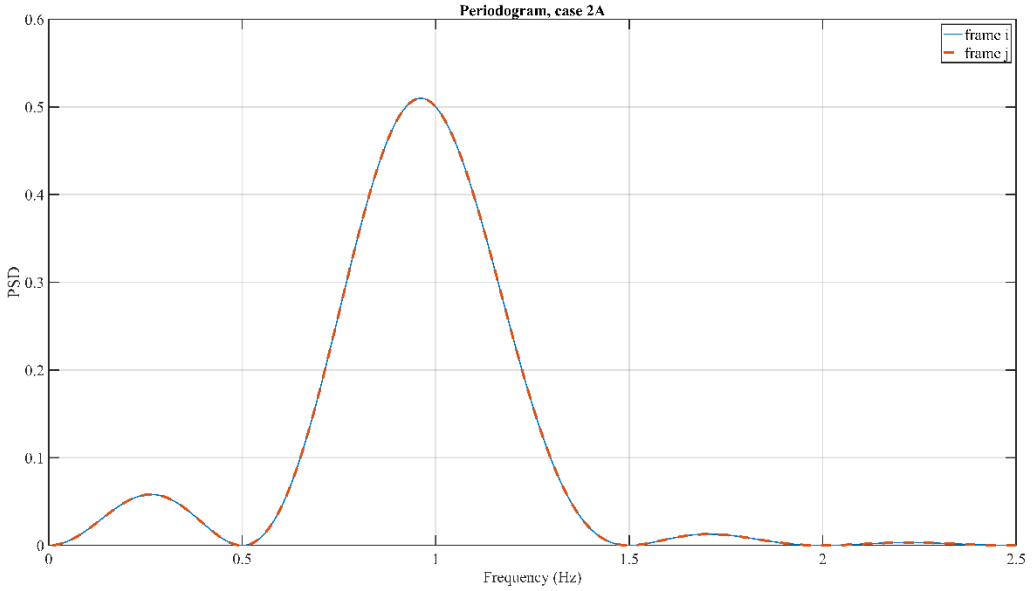
Case $(2A)$: Frame $j$ is replayed over frame $i$ and the replay of frame $j$ covers all of frame $i$.

Let $\alpha$ be the portion of the frame $j$ replayed over the length of frame $i$. Therefore, in case $(2A)$:

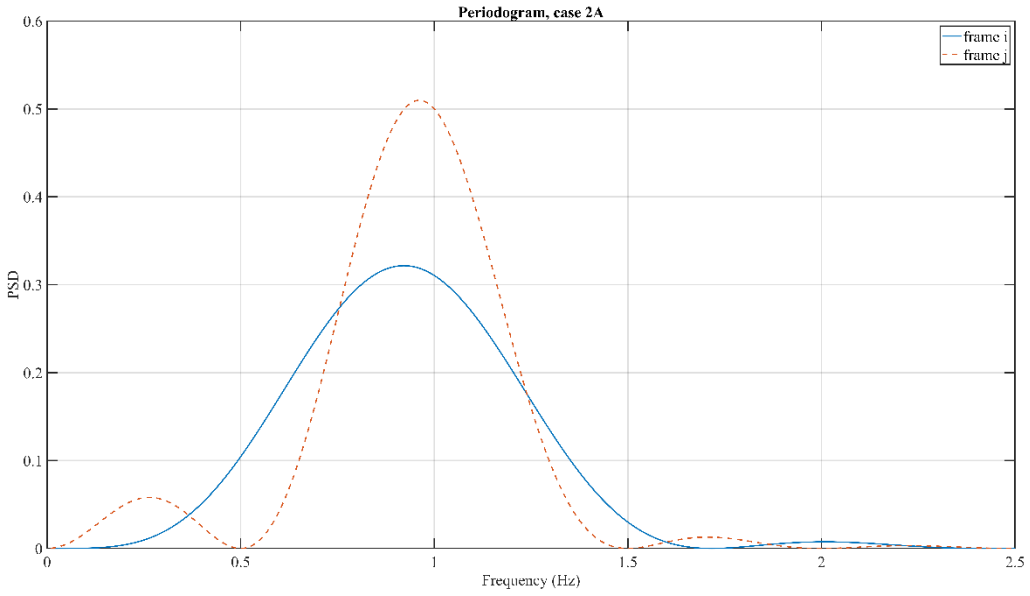
$$\alpha = \frac{T_{f_i}}{T_{f_j}} \qquad (3.28)$$

Since $T_{f_j} = kT_{comb}$ and $k \geq 2$, for each frequency in frame $j$, at least one period is played back given the assumption $T_{f_i} < 2T_{f_j}$.

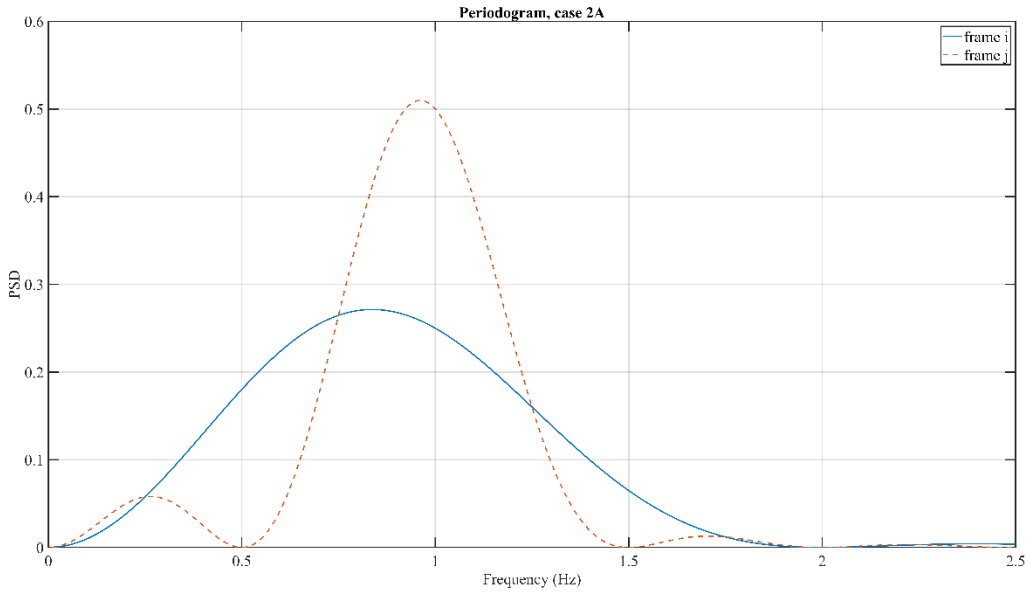
*Example 3.2.* Consider the watermark to be $x(t) = \sin(2\pi f_j t)$ where $f_j = 1\,Hz, T_j = 1s, k = 2$ and $T_{f_j} = 2s$ that is replayed on frame $i$ with $T_s = 0.001s$. For different values of $\alpha$, the periodogram of frame $i$ compared to the original periodogram of frame $j$ is shown in $Fig.3.12$.




$(a)$




$(b)$




$(c)$

*Fig.* 3.12. Periodogram of frame $i$ during attack for $(a)$ $\alpha = 1, T_{f_i} = 2s$ $(b)$ $\alpha = 0.75, T_{f_i} = 1.5s$ $(c)$ $\alpha = 0.5, T_{f_i} = 1s$.

We observe that as the portion of frame $j$ replayed, $\alpha$, decreases, the PSD widens making it harder to detect the frequencies present in the replayed data using a periodogram. Therefore, we can conclude that the periodogram of frame $i$ will be similar to that of frame $j$ only with a resolution

of $\frac{1}{T_{f_j}}$ worsened to $\frac{1}{T_{f_i}}$. So, the PSD will spread out when compared to the original form of periodogram of frame $j$.

Case $(2B)$: Frame $j$ and frame $j+1$ are replayed over frame $i$. Frequencies $f_1^j, \dots f_{n_m}^j, f_1^{j+1}, \dots f_{n_m}^{j+1}$ are present in frame $i$ under attack.

We consider that $\alpha_1$ portion of frame $j$ is replayed in frame $i$ and $\alpha_2$ portion of frame $j+1$ is replayed for the rest of frame $i$. Therefore, the signal replayed over frame $i$ is

$$x(t) = \begin{cases} x_j(t - t_0) & 0 < t \leq \alpha_1 T_{f_j} \\ x_{j+1}(t - t_0) & \alpha_1 T_{f_j} < t \leq T_{f_i} \end{cases} \tag{3.29}$$

where $t_0$ is the time from when the attacker records frame $j$ and to the time it replays frames $j$ and $j+1$.

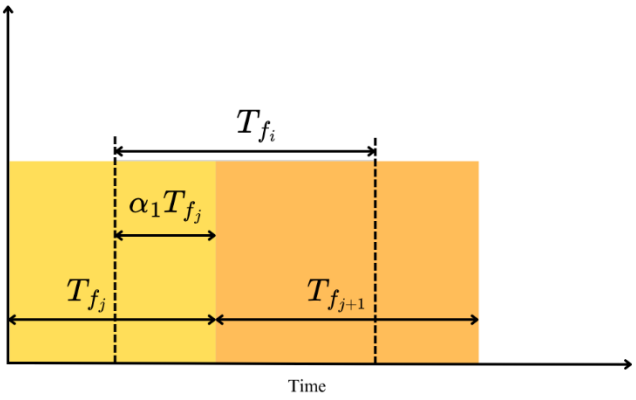



*Fig.* 3.13. Significance of nomenclature used.

Let $\beta$ be the portion of the frame $i$ that contains replayed frame $j$, then, $\beta T_{f_i} = \alpha_1 T_{f_j}$. The replayed signal can be rewritten as

$$x(t) = \begin{cases} x_j(t - t_0) & 0 < t \leq \beta T_{f_i} \\ x_{j+1}(t - t_0) & \beta T_{f_i} < t \leq T_{f_i} \end{cases} \tag{3.30}$$

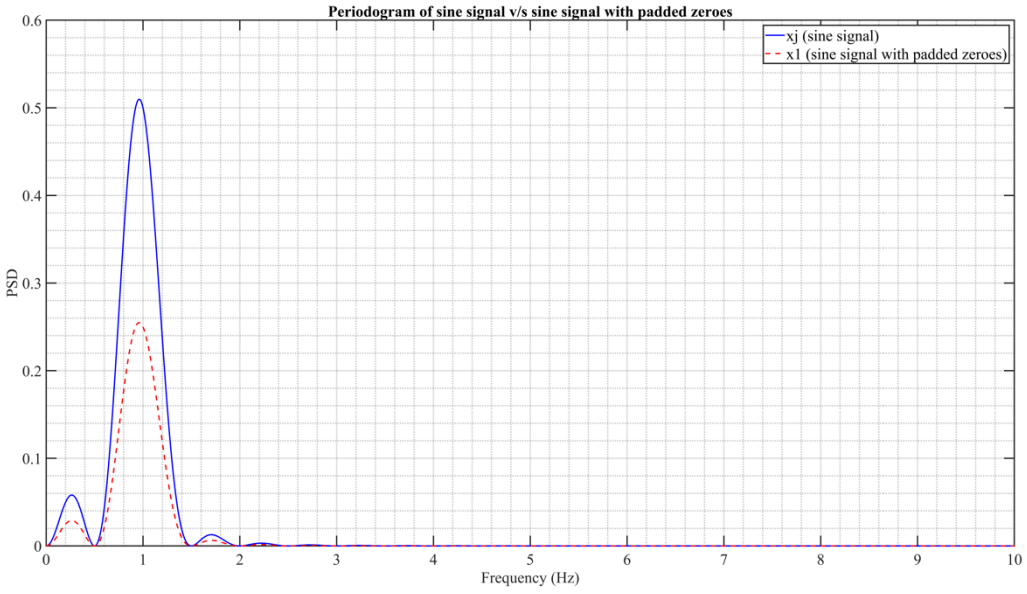
Let $x_1(t)$ and $x_2(t)$ denote the above two segments padded with zeroes:

$$x_1(t) = \begin{cases} x_j(t - t_0) & 0 < t \leq \beta T_{f_i} \\ 0 & \beta T_{f_i} < t \leq T_{f_i} \end{cases}$$

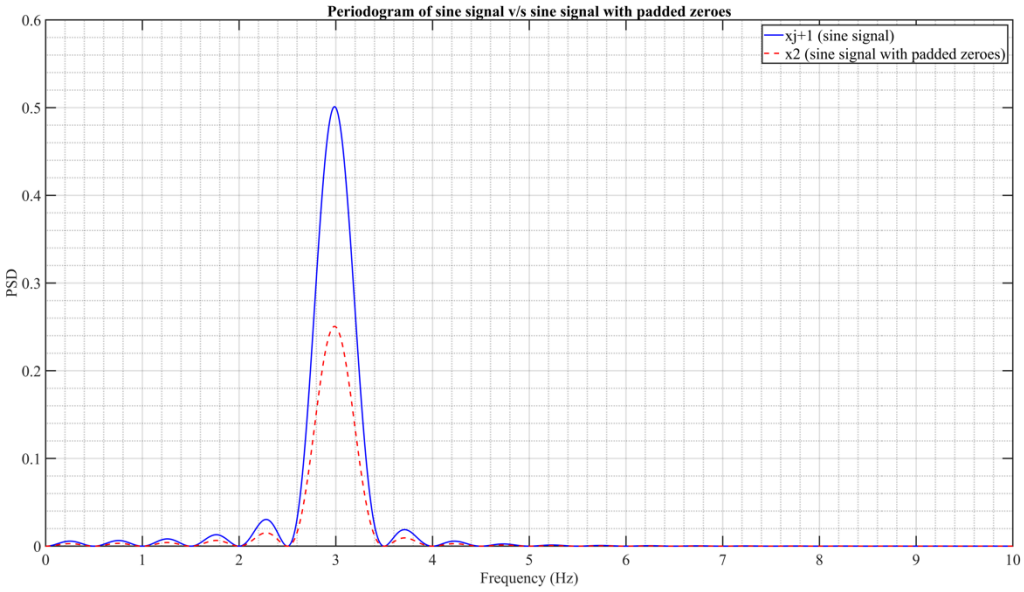
$$x_2(t) = \begin{cases} 0 & 0 < t \leq \beta T_{f_i} \\ x_{j+1}(t - t_0) & \beta T_{f_i} < t \leq T_{f_i} \end{cases}$$

Based on the conclusion drawn in section 2.4 on padding a signal with zero, we can observe that the periodogram of the samples of $x_j(t - t_0)$ over $0 \leq t \leq \beta T_{f_i}$ and the periodogram of samples of $x_1(t)$ over $0 \leq t \leq T_{f_i}$ are the same except that the latter is a scaled version of the former by a factor of $\beta$. Similarly, the periodogram of $x_2(t)$ is $1 - \beta$ times the periodogram of samples of $x_{j+1}(t - t_0)$ over $\beta T_{f_i} \leq t \leq T_{f_i}$.

*Example 3.3.* Assume the watermark for frames $j$ and $j + 1$ to be $x_j(t) = \sin(2\pi f_j t)$ where $f_j = 1\ Hz$ and $x_{j+1}(t) = \sin(2\pi f_{j+1} t)$ where $f_{j+1} = 3\ Hz$, respectively, replayed over frame $i$ with $T_{f_i} = 4s$, $\beta = 0.5$ and sampling frequency $f_s = 1000\ Hz$. The periodograms of $x_j(t)$ v/s $x_1(t)$ and $x_{j+1}(t)$ v/s $x_2(t)$ are presented in $Fig.3.14\ (a)$ and $(b)$, respectively.



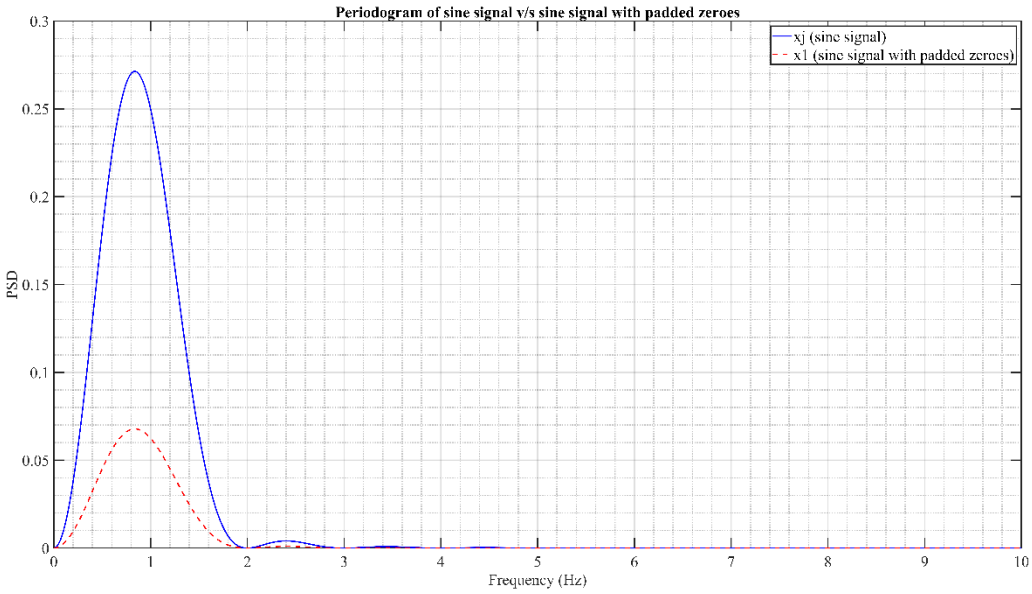

$(a)$



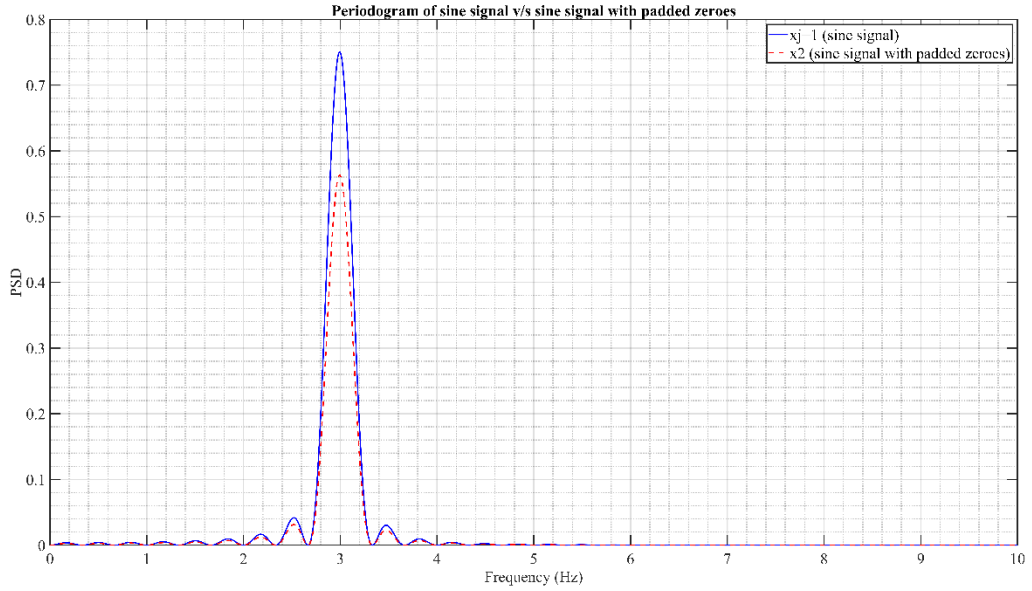

$(b)$

$Fig.3.14$. *Example 3.3.* Periodogram of sine signal v/s sine signal with padded zeroes in $(a)$ frame $j$ and $(b)$ frame $j+1$ with $\beta = 0.5$.

From $Fig.3.14$, we observe that the peak amplitude of the periodogram is multiplied by a factor of $0.5$ in frames $j$ and $j+1$. The same example is simulated for a different value of $\beta$, namely, $\beta = 0.25$ and the results are presented in $Fig.3.15$.




$(a)$




$(b)$

*Fig.* 3.15. *Example 3.3.* Periodogram of sine signal v/s sine signal with padded zeroes in $(a)$ frame $j$ and $(b)$ frame $j+1$ with $\beta = 0.25$.

From *Fig.* 3.15, we verify that the peak amplitude of periodograms of a sine signals $x_j$ and $x_{j+1}$ padded with zeroes reduce by a factor of 0.25 and 0.75, respectively.

*Proposition 1.* The PSD of $x(t)$ is the sum of PSD of $x_1(t)$ and PSD of $x_2(t)$.

*Proof.* We know that

$$x(t) = x_1(t) + x_2(t)$$

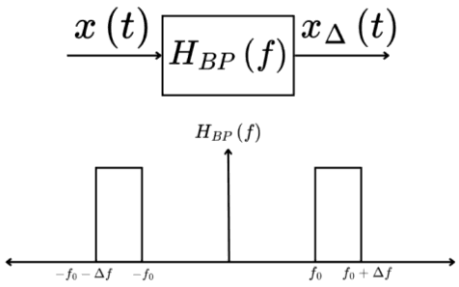
and energy over $[0, T_{f_i}]$

$$\int_0^{T_{f_i}} |x(t)|^2 dt = \int_0^{T_{f_i}} |x_1(t)|^2 dt + \int_0^{T_{f_i}} |x_2(t)|^2 dt \tag{3.31}$$

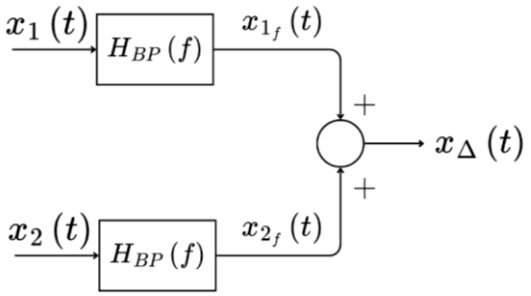
since $x_1(t)x_2(t) = 0$ for $0 \leq t \leq T_{f_i}$. Next, we show the energy in small frequency range $[f_0, f_0 + \Delta f]$ in $x(t)$ is the sum of energies in $x_1(t)$ and $x_2(t)$ in that frequency range. The energy in $[f_0, f_0 + \Delta f]$ can be obtained from $x_\Delta(t)$ as shown in $Fig. 3.16(a)$. Here, $x_\Delta(t)$ is $x(t)$ filtered by an ideal band pass filter (BPF) $H_{BP}(f)$. Due to the linearity of BPF,

$$x_\Delta(t) = x(t) * h_{BP}(t) = x_1(t) * h_{BP}(t) + x_2(t) * h_{BP}(t) \tag{3.32}$$

i.e. passing $x(t)$ through the BPF gives the same result as passing $x_1(t)$ and $x_2(t)$ separately through the BPF and then summing the outputs as shown in $Fig. 3.16(b)$. Therefore, the energy in $x_\Delta(t)$ is the sum of energies in $x_{1_f}(t)$ and $x_{2_f}(t)$. From this the proposition follows.




$(a)$




$(b)$

$Fig.$ 3.16. Signal $(a)$ $x(t)$ passed through the BPF and $(b)$ $x_1(t)$ and $x_2(t)$ passed through the BPF and summed.

Despite the proposition, the periodogram of $x(t)$ is not necessarily the sum of the periodograms of $x_1(t)$ and $x_2(t)$:

$$P_{xx}(f) \neq P_{x_1x_1}(f) + P_{x_2x_2}(f)$$

This can be justified by the fact that periodogram is an estimate for PSD. We explore this issue further using two specific cases.

(a) Suppose $x_1(t)$ and $x_2(t)$ are sinusoids with same frequencies.

$$x_1(t) = \begin{cases} \sin \omega_0 t & 0 \leq t \leq T_0 \\ 0 & T_0 \leq t \leq 2T_0 \end{cases}$$

$$x_2(t) = \begin{cases} 0 & 0 \leq t \leq T_0 \\ \sin \omega_0 (t - T_0) & T_0 \leq t \leq 2T_0 \end{cases}$$

where $T_0 = \frac{2\pi}{\omega_0}$, $T_{f_i} = 2T_0$ and $T_0 = NT_s$ where $T_s$ is the sampling time. Hence $T_{f_i} = 2NT_s$.

$$\begin{aligned} x_2[n] &= x_2(nT_s) = x_1\big(\omega_0(nT_s - T_0)\big) \qquad N \leq n \leq 2N-1 \\ &= x_1\big(\omega_0(n-N)T_s\big) \\ &= x_1[n-N] \end{aligned}$$

Therefore, for discrete-time Fourier Transform:

$$X_2\big(e^{j\Omega}\big) = X_1\big(e^{j\Omega}\big)e^{-j\Omega N}$$

$$|X(e^{j\Omega})|^2 = |X_1(e^{j\Omega}) + X_2(e^{j\Omega})|^2$$

$$= X_1(e^{j\Omega})X_1^*(e^{j\Omega}) + X_1(e^{j\Omega})X_2^*(e^{j\Omega}) + X_1^*(e^{j\Omega})X_2(e^{j\Omega}) + X_2(e^{j\Omega})X_2^*(e^{j\Omega})$$

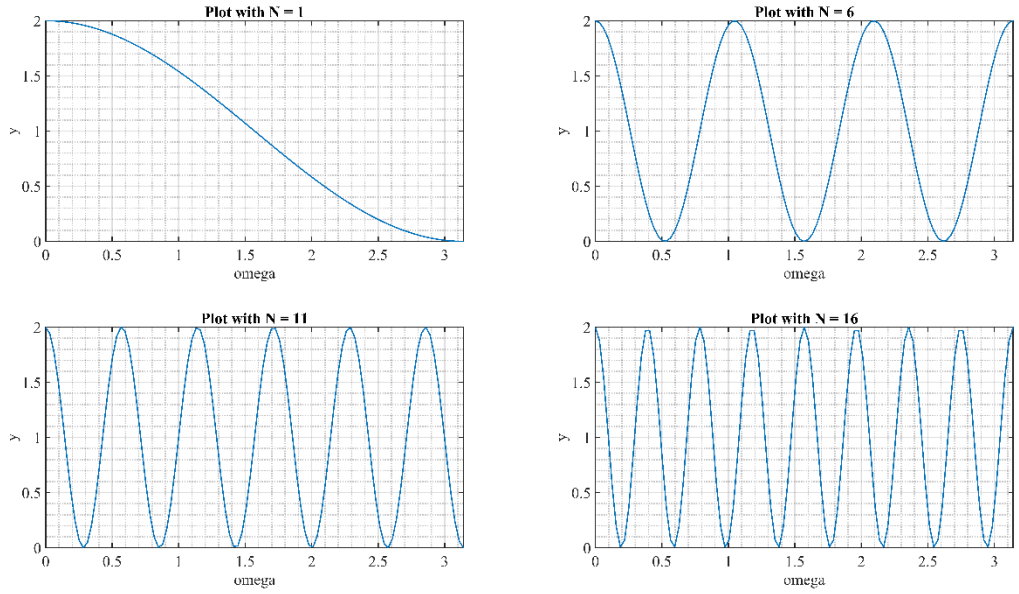
$$= |X_1(e^{j\Omega})|^2 + |X_1(e^{j\Omega})|^2 e^{j\Omega N} + |X_1(e^{j\Omega})|^2 e^{-j\Omega N} + |X_1(e^{j\Omega})|^2$$

$$= 2|X_1(e^{j\Omega})|^2(1 + \cos N\Omega)$$

That means

$$P_{xx}(f) = 2P_{x_1x_1}(f)(1 + \cos N2\pi f T_s) \neq P_{x_1x_1}(f) + P_{x_2x_2}(f) \qquad (3.33)$$

In $Fig.3.17$, function $(1 + \cos N\Omega)$ is plotted for $0 \leq \Omega \leq \pi$.




$Fig.3.17.$ $(1 + \cos N\Omega)$ plotted for $0 \leq \Omega \leq \pi$.

(b) Suppose $x_1(t)$ and $x_2(t)$ are sinusoids with different frequencies.

$$x_1(t) = \sin \omega_1 t$$

$$x_2(t) = \sin \omega_2 t$$

and assume

$$\omega_2 - \omega_1 = \frac{f_2 - f_1}{2\pi} \gg \Delta f_i = \frac{1}{T_{f_i}}$$

i.e. frequencies are far apart. Then, the energy of $x_1$ is concentrated around $f_1 = \frac{\omega_1}{2\pi}\ Hz$ and the energy of $x_2$ around $f_2 = \frac{\omega_2}{2\pi}\ Hz$. Hence, $X_1(e^{j\Omega})X_2(e^{j\Omega}) \sim 0$ for any $\Omega$ because either $X_1(e^{j\Omega}) = 0$ or $X_2(e^{j\Omega}) = 0$.

$$\left|X(e^{j\Omega})\right|^2 = \left|X_1(e^{j\Omega}) + X_2(e^{j\Omega})\right|^2$$

$$= X_1(e^{j\Omega})X_1^*(e^{j\Omega}) + X_1(e^{j\Omega})X_2^*(e^{j\Omega}) + X_1^*(e^{j\Omega})X_2(e^{j\Omega}) + X_2(e^{j\Omega})X_2^*(e^{j\Omega})$$

$$\cong \left|X_1(e^{j\Omega})\right|^2 + \left|X_2(e^{j\Omega})\right|^2$$

Therefore,

$$P_{xx}(f) \cong P_{x_1x_1}(f) + P_{x_2x_2}(f) \tag{3.34}$$

where $P_{xx}$, $P_{x_1x_1}$ and $P_{x_2x_2}$ represent the periodograms of signals $x(t)$, $x_1(t)$ and $x_2(t)$, respectively.

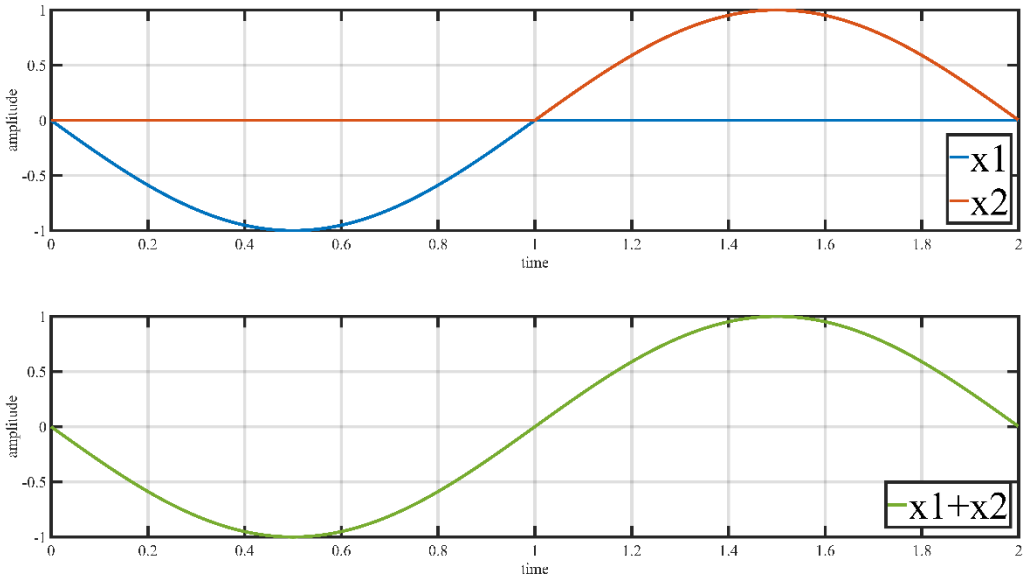
*Example 3.4.* Consider the same replay signal as presented in *Example 3.3*. We have

$$x_1(t) = \begin{cases} \sin(2\pi f_j t) & 0 < t \leq \beta T_{f_i} \\ 0 & \beta T_{f_i} < t \leq T_{f_i} \end{cases}$$

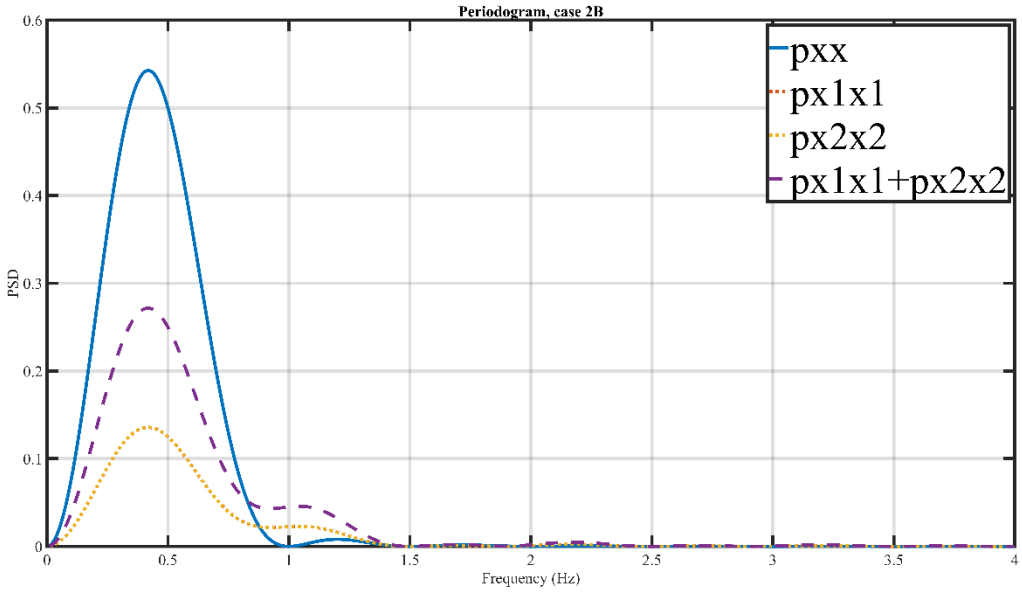
$$x_2(t) = \begin{cases} 0 & 0 < t \leq \beta T_{f_i} \\ \sin(2\pi f_{j+1} t) & \beta T_{f_i} < t \leq T_{f_i} \end{cases}$$

This time $T_{f_i}$ is shorter with $T_{f_i} = 2s$ and the corresponding frequency resolution of $\Delta f_i = \frac{1}{T_{f_i}} = 0.5\ Hz$. With sampling time $T_s = 0.001s$ and $\beta = 0.5$, we consider the following four cases.

(i) The first case is an example of case (a). Let $f_j = f_{j+1} = 0.5\ Hz$, $T_{f_j} = T_{f_{j+1}} = 4s$, $\alpha_1 = 0.25$ and $T_{f_i} = 2s$. The time domain representation of these signals followed by the comparison of periodogram of these signals over $[0, T_{f_i}]$ are shown in $Fig.\ 3.18$.




$(a)$

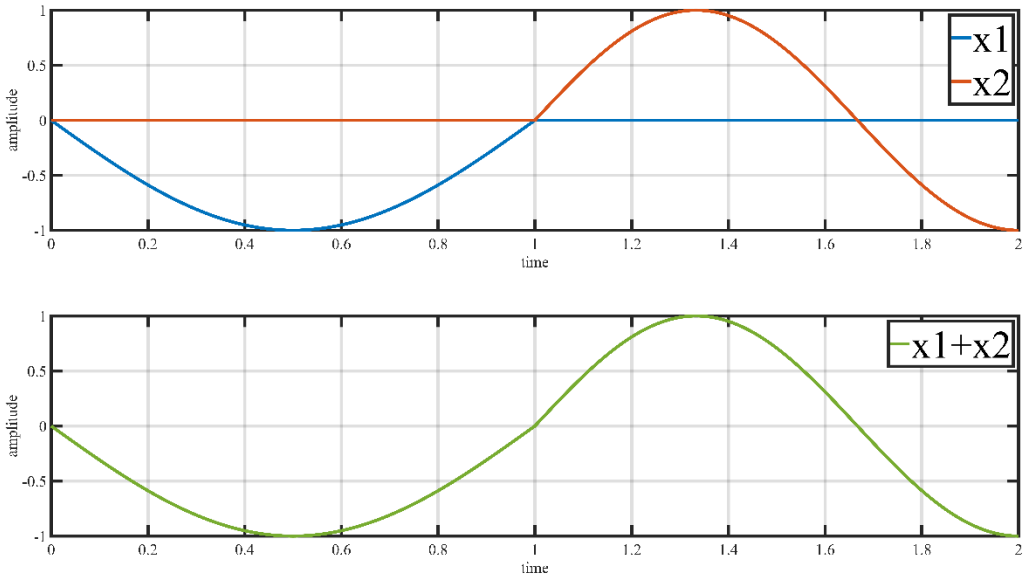



$(b)$

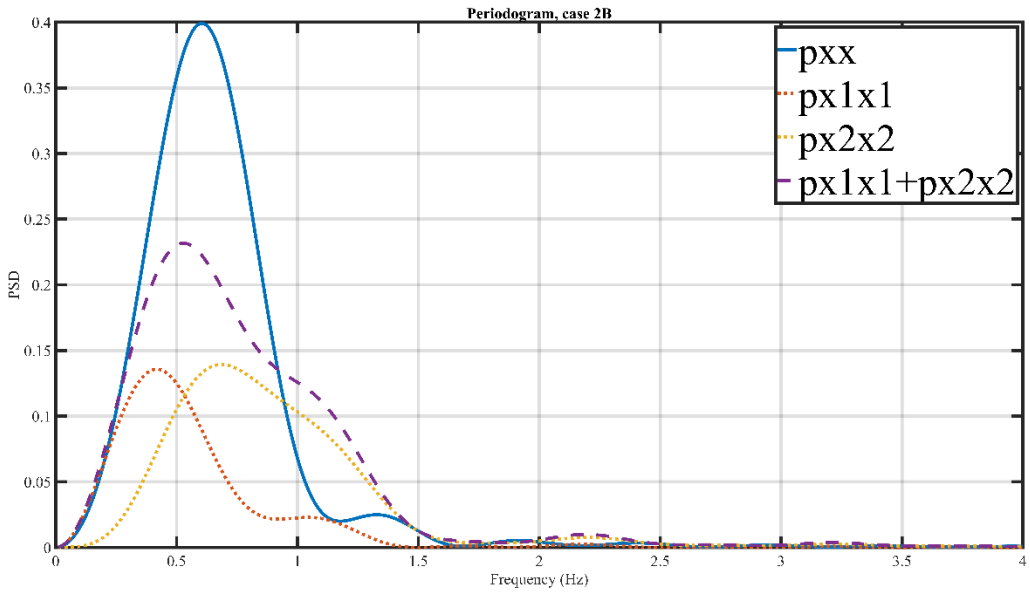
*Fig.* 3.18. *Example 3.4(i)* $(a)$ Time domain representation of $x_1(t)$ and $x_2(t)$ $(b)$ Periodogram of $x_1(t)$, $x_2(t)$ and $x(t)$ on the same scale.

From *Fig.* 3.18, we can observe that $P_{xx} \neq P_{x_1x_1} + P_{x_2x_2}$

(ii) Next, consider different but close frequencies. Let $f_j = 0.5\ Hz$ , $f_{j+1} = 0.75\ Hz, T_{f_j} = T_{f_{j+1}} = 4s, \alpha_1 = 0.25, f_i = 0.5$ and $T_{f_i} = 2s$. The time domain representation of these signals followed by the comparison of periodograms of these signals over $[0, T_{f_i}]$ are shown in $Fig.3.19$.




$(a)$

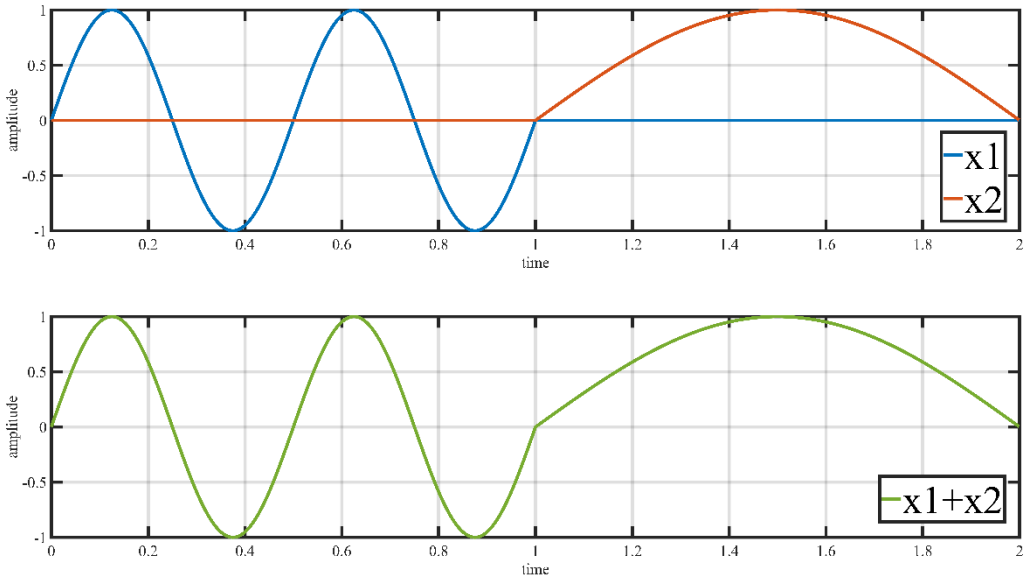



$(b)$

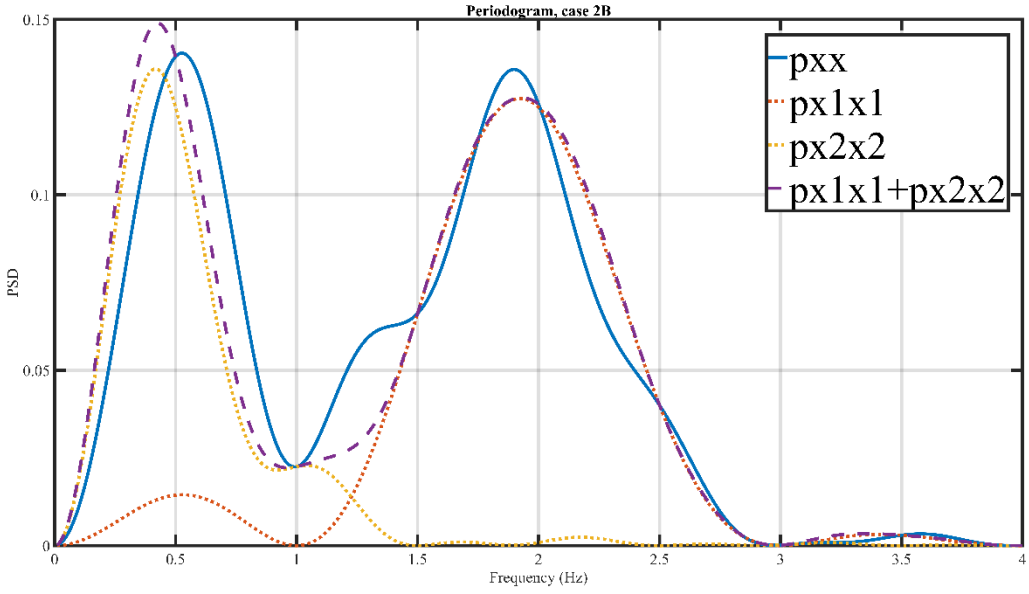
$Fig.3.19$. *Example 3.4(ii)* $(a)$ Time domain representation of $x_1(t)$ and $x_2(t)$ $(b)$ Periodogram of $x_1(t)$, $x_2(t)$ and $x(t)$ on the same scale.

From $Fig.3.19$, we can observe that $P_{xx}$ and $P_{x_1x_1} + P_{x_2x_2}$ are very different and the frequencies of $x_1$ and $x_2$ (0.5 $Hz$ and 0.75 $Hz$) are not visible in $P_{xx}$.

(iii) Let $f_j = 2\ Hz$, $f_{j+1} = 0.5\ Hz$, $T_{f_j} = T_{f_{j+1}} = 4s$, $\alpha_1 = 0.25$, $f_i = 1\ Hz$ and $T_{f_i} = 2s$. The time domain representation of these signals followed by the comparison of periodograms of these signals over $[0, T_{f_i}]$ are shown in $Fig.3.20$.




$(a)$




$(b)$

*Fig.* 3.20. *Example 3.4(iii)* $(a)$ Time domain representation of $x_1(t)$ and $x_2(t)$ $(b)$ Periodogram of $x_1(t)$, $x_2(t)$ and $x(t)$ on the same scale.

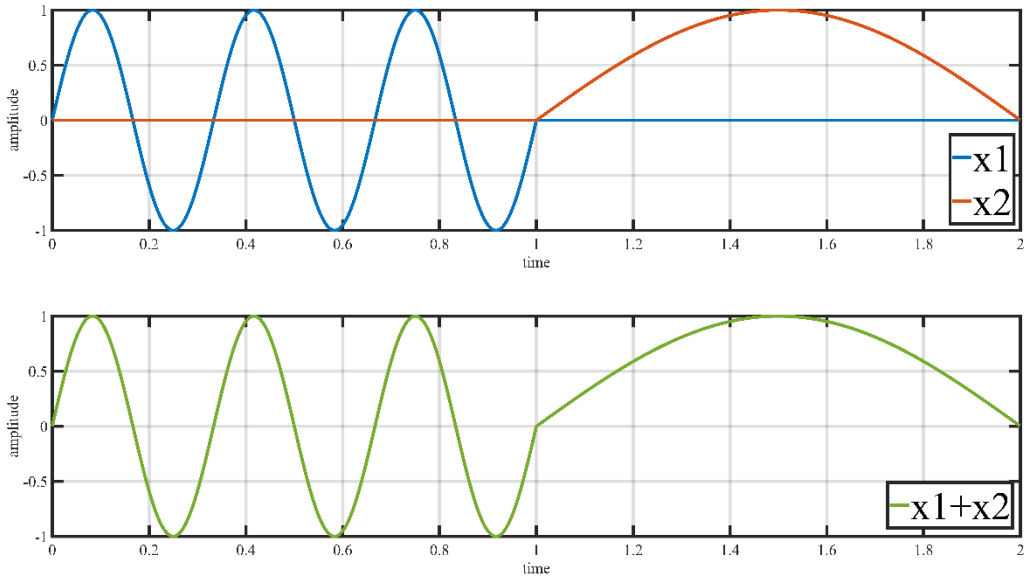
From *Fig.* 3.20, we can observe that frequencies of $x_1$ and $x_2$ can be seen in $P_{xx}$ but a large side lobe at 1.25 $Hz$ is also present. In this case,

$$f_j - f_{j+1} = 1.5\ Hz = 3\Delta f_i$$

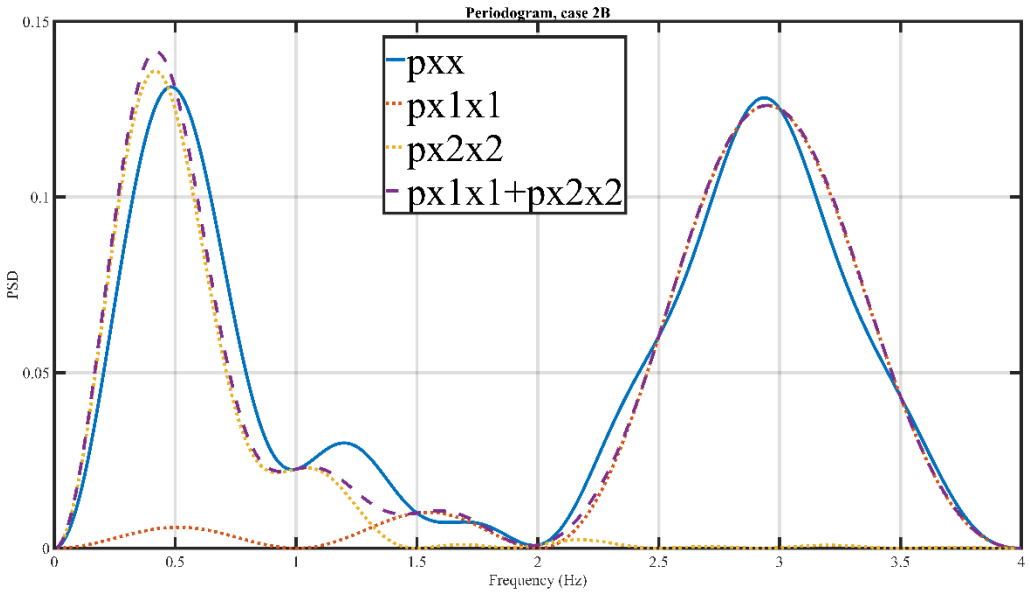
(iv) Let $f_j = 3\ Hz$, $f_{j+1} = 0.5\ Hz$, $T_{f_j} = T_{f_{j+1}} = 4s$ and $\alpha_1 = 0.25$. The time domain representation of these signals followed by the comparison of periodogram of these signals over $[0, T_{f_i}]$ is shown in $Fig.\ 3.21$. In this case,

$$f_j - f_{j+1} = 2.5\ Hz = 5\Delta f_i$$

We can see that the two peaks at frequencies $f_j = 3\ Hz$ and $f_{j+1} = 0.5\ Hz$ can be observed and $P_{xx} \cong P_{x_1x_1} + P_{x_2x_2}$.




$(a)$




$(b)$

*Fig.* 3.21. *Example 3.4(iv)* $(a)$ Time domain representation of $x_1(t)$ and $x_2(t)$ $(b)$ Periodogram of $x_1(t)$, $x_2(t)$ and $x(t)$ on the same scale.

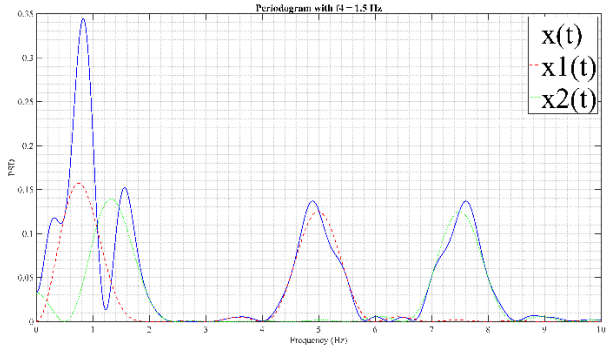
To summarize, from *Fig.* 3.21, we can observe that frequencies of $x_1$ and $x_2$ are distinguishable in $P_{xx}$.

When the frequencies $f_j$ and $f_{j+1}$ are close to each other, the periodogram is deformed and does not yield the clean peaks at frequencies as expected due to spectral leakage. However, when the frequencies are far apart, the periodogram holds the expected shape with clean peaks letting us detect the frequencies present. Therefore, when placed sufficiently apart, we can use $P_{xx}(f) = P_{x_1x_1}(f) + P_{x_2x_2}(f)$ in all aspects of our analysis. Hence, we conclude that the separation of the frequencies should be at least $4\Delta f_i$ for (3.32) to tend to equality. This further yields the result that in a replay attack of the type considered in case $(2B)$, the frequencies present in the replayed signal are successfully detected when found to be sufficiently apart as in *Fig.* 3.21$(b)$. In *Fig.* 3.19$(b)$, since the frequencies are close to each other, the periodogram cannot detect the frequencies present. In either case, the presence of unexpected frequencies or the absence of expected frequencies is confirmed during detection, thereby conveying the occurrence of a replay attack. However, in this conclusion there was only one frequency present per frame. The same result is investigated for two frequencies present per frame in the following example.

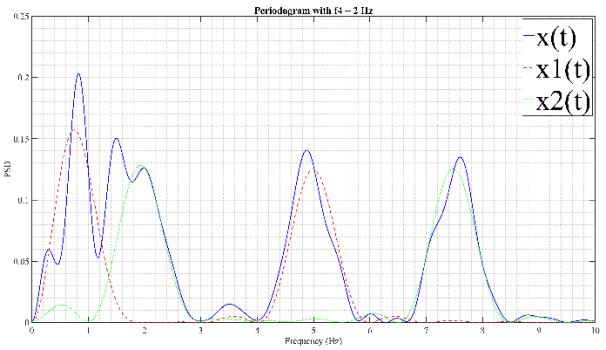
*Example 3.5.* Consider the replayed signal to be

$$x(t) = \begin{cases} x_1(t) & 0 < t \leq \beta T_{f_i} \\ x_2(t) & \beta T_{f_i} < t \leq T_{f_i} \end{cases}$$

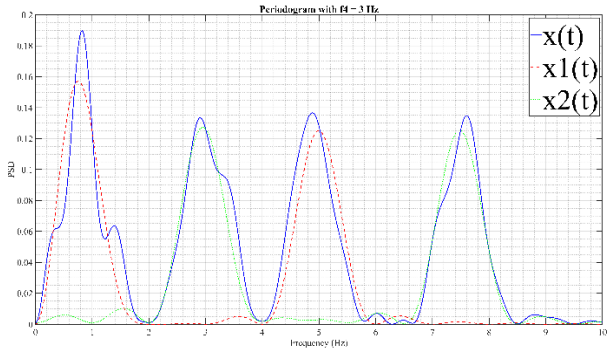
where $x_1(t) = \sin(2\pi f_1 t) + \sin(2\pi f_2 t)$ where $f_1 = 1\,Hz$ and $f_2 = 5\,Hz$ and $x_2(t) = \sin(2\pi f_3 t) + \sin(2\pi f_4 t)$ where $f_3 = 7.5\,Hz$ with $\beta = 0.5$ replayed over frame $i$ with $f_i = 0.5$ and $T_{f_i} = 2\,s$. With resolution, $\Delta f_i = \frac{1}{T_{f_i}} = 0.5\,Hz$ and sampling period, $T_s = 0.001s$, multiple values of $f_4 = [1.5, 2, 3, 4, 4.5, 5.5, 6, 7, 8, 8.5]\,Hz$ are considered to demonstrate the complexity in detection of multiple frequencies present in frames with reduced resolution as shown in $Fig.3.22$.



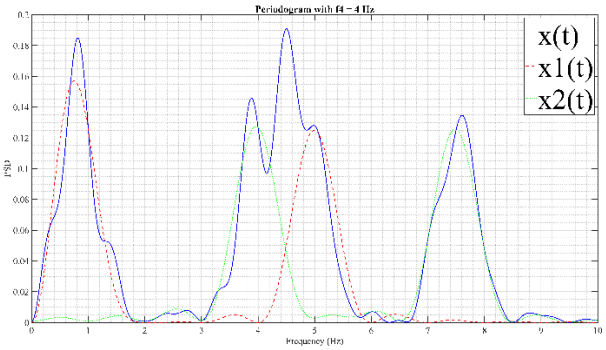

$(a)\ f_4 = 1.5\,Hz$



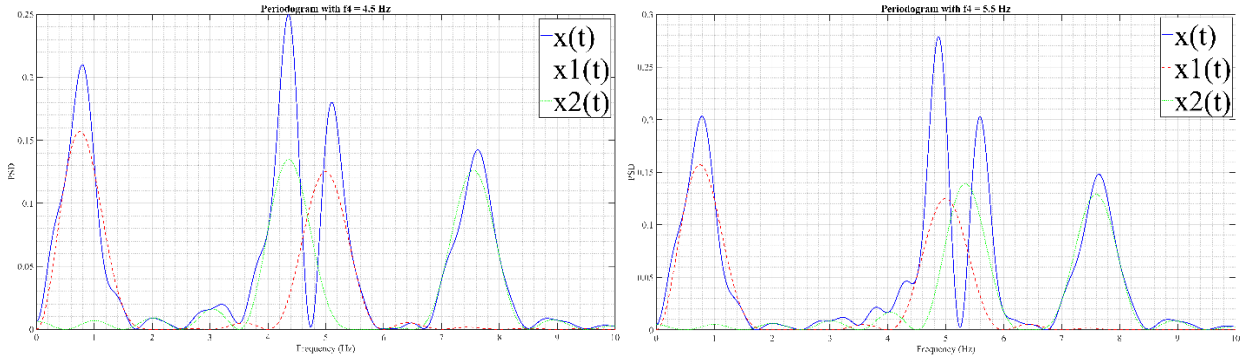

$(b)\ f_4 = 2\,Hz$




$(c)\ f_4 = 3\,Hz$



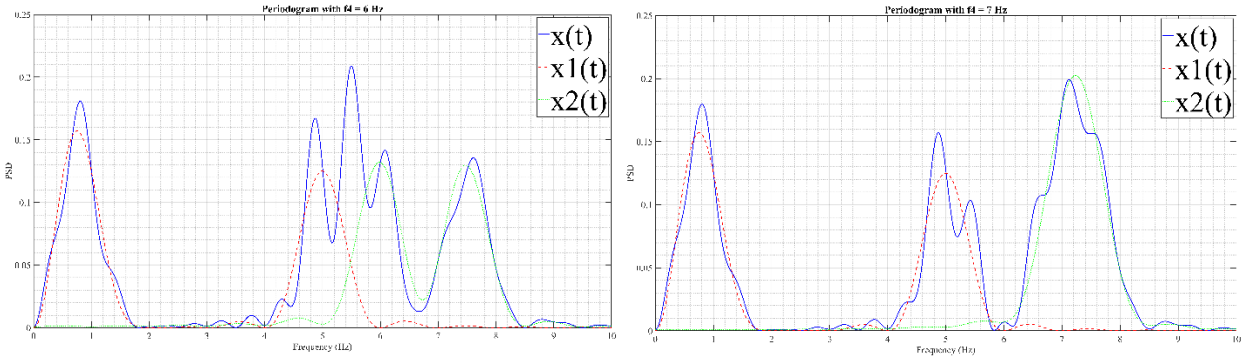

$(d)\ f_4 = 4\,Hz$




$(e)\ f_4 = 4.5\ Hz$



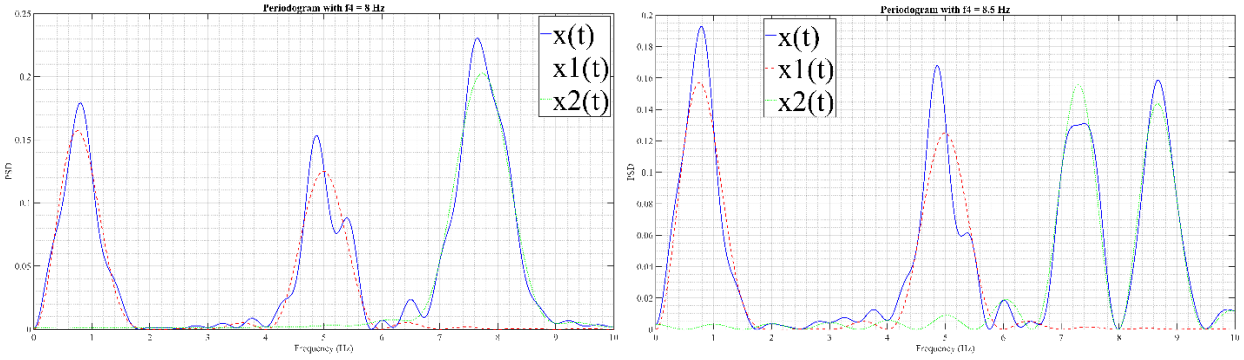

$(f)\ f_4 = 5.5\ Hz$




$(g)\ f_4 = 6\ Hz$




$(h)\ f_4 = 7\ Hz$




$(i)\ f_4 = 8\ Hz$




$(j)\ f_4 = 8.5\ Hz$

*Fig.* 3.22. Periodogram of $x_1(t)$, $x_2(t)$ and $x(t)$ on the same scale.

When the frequencies get too close to each other, i.e. when any two of the four frequencies are $\sim\Delta f_i = 0.5\ Hz$ apart from each other (shown in $Fig.\ 3.22(a), (e), (f), (h), (i)$), the PSD lobes

interact with each other and their magnitude increases compared with the other peaks observed in the periodogram. Therefore, in these cases, the periodogram detects inaccurate frequencies (i.e. frequencies that were not present in frame $j$ or $j+1$) to be present in the system. This can also be viewed as a complex extension of case (ii) of *Example 3.4*.

When frequencies are $\sim 2\Delta f_i = 1\ Hz$ apart from each other (shown in $Fig.\ 3.22(b), (d), (g), (j)$), the individual PSD lobes interact with each other and spread out the resulting PSD due to the interaction of the side lobes. In these cases, the periodogram is unable to detect conclusively frequencies of frame $j$ or $j+1$. This case is an extension of case (iii) of *Example 3.4*.

When frequencies are $\sim 4\Delta f_i = 2\ Hz$ apart from each other (shown in $Fig.\ 3.22(c)$), the individual PSD peaks are clear and the periodogram is able to detect frequencies of frame $j$ or $j+1$. This case can be considered an extension of case (iv) of *Example 3.4*.

Hence, the results obtained from *Example 3.4* can be extended for more complex cases as necessary. However, a change in the value of $\beta$ will result in an unequal presence of frames in the replayed signal ending up suppressing the frequencies of the shorter frame. For instance, if $\beta = 0.25$, then frame $j$ will be shorter in the replayed signal and the frequencies of frame $j$ will have a lower PSD value in the periodogram. In any case, if the frequencies are placed apart enough, the detection of frequencies of frame $j+1$ will point us towards the presence of a replay attack.

Case $(2C)$ is the scenario where the replayed data is from three different frames. We analyze the possible outcomes based on the proportion of signal replayed from each frame. Let $t_1$ be the duration of the first frame replayed in the $i^{th}$ frame, $t_2$ the duration of the second frame replayed in the $i^{th}$ frame and $t_3$ the duration of the third frame replayed in the $i^{th}$ frame as shown in $Fig.\ 3.23$.

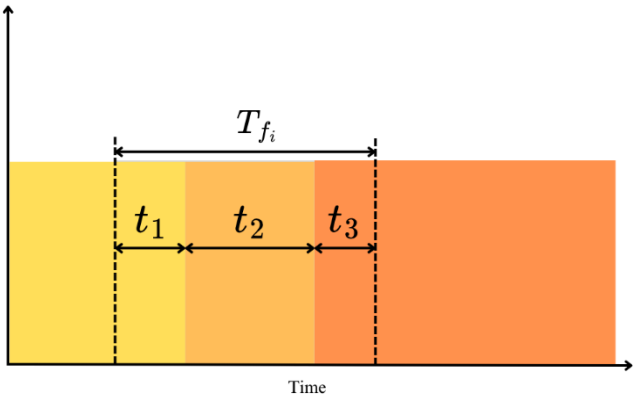



*Fig*. 3.23. Depiction of Case $(2C)$.

Let $x(t)$ denote the output of frame $i$ under replay attack and suppose it is a combination of sinusoidal signals described as:

$$x(t) = \begin{cases} \sum_{i=1}^{n_m} A_i^1 \sin(\omega_i^1 t + \phi_i^1) & t_0 < t \leq (t_0 + t_1) \\ \sum_{i=1}^{n_m} A_i^2 \sin(\omega_i^2 t + \phi_i^2) & (t_0 + t_1) < t \leq (t_0 + t_1 + t_2) \\ \sum_{i=1}^{n_m} A_i^3 \sin(\omega_i^3 t + \phi_i^3) & (t_0 + t_1 + t_2) < t \leq (t_0 + t_1 + t_2 + t_3) \end{cases}$$

where $t_1 + t_2 + t_3 = T_f^i$. When $t_1, t_3 \ll t_2$, this case can be treated as case $(2A)$ with $\alpha = 1$ since we can approximate $t_2 \sim T_{f_i}$ and the replayed data can be considered to be from the second frame for the entire duration. Also, case $(2C)$ can be considered as an extension of case $(2B)$ with three different frames present in the replayed signal and a similar conclusion can be extended to this case. Namely, if a frame has a sinusoid with a frequency $f_0$ and $f_0$ is sufficiently apart from sinusoids, then $f_0$ will be detectable in $P_{xx}(f)$.

## 3.4 Proposed Watermark Design Guidelines

In this section, based on the analysis of the previous section, we present guidelines for choosing the parameters of the multi-sine watermark. We begin by providing a summary of the conclusions of the previous section.

Case 1 ($\delta \ll T_{f_i}$ for $i = 1,2, \dots, n_f$): PSD contains frequency $\frac{2\pi}{\delta}$ and its harmonics.

Case 2 ($\delta \gg T_{f_i}$ for $i = 1,2, \dots, n_f$) divided into the following sub-cases:

Case $2A$: PSD includes frequencies of frame $j$. In the extreme case $\alpha = 0.5$, the PSD drops in magnitude by a factor of $0.5$ or reduces by $3\ dB$.

Case $2B$: The frequency of frame $j$ can be observed in PSD if it is separated from other frequencies by at least $4\Delta f_i$. In this case, the periodogram with confidence interval for the worst case $\alpha = 0.5$, $\beta = 0.5$ gives the false negative rate.

Case $2C$: This case is considered an extension of Case $2B$. The conclusion of Case $2B$ can be extended to this case.

If each frame has a unique frequency at least $4\Delta f_i$ apart from other frequencies in that frame and all other frames, then we can detect the presence of the frequency during attack with the desired confidence interval.

This confidence interval gives the false negative rate. For case $(2A)$, the worst case for detecting the replay attack is when $\alpha = 0.5$. In this case, the periodogram at the identifying frequency drops by $0.5$ (or $-3dB$). The gain $c_0$ in watermark can be chosen to have sufficient confidence for case $(2B)$. For cases $(2B)$ and $(2C)$, a similar argument holds, with the worst case for detection occurring when $\alpha = 0.5$ and $\beta = 0.5$.

Therefore, in designing the multi-sine watermarking signal, the range of frequencies $f_1^i, \dots, f_{n_m}^i$ must be set in a strategic manner (Note that the number of frequencies per frame $n_m$ is chosen from $(3.8)$). One approach is described in the following.

Each frequency is placed within certain frequency intervals we call *bins*. These frequency bins are at least $\max 4\Delta f_i = \max \frac{4}{T_{f_i}}$ apart from each other. In order to be able to detect replay attack, we make sure that each frame has at least one unique frequency with its bin. No other frequency will be in that bin. Let $n_b$ be the number of frequency bins and

$$\Delta f_b = 4 * \max(\Delta f_i) \qquad \text{for } i = 1,2, \dots, n_f \qquad (3.35)$$

where $n_f$ is the number of frames. Then, to satisfy the above conditions:

$$n_b \geq n_f + n_m - 1 \tag{3.36}$$

where $2 \leq n_f \leq n_b$. The gain $c_0$ is chosen to provide sufficiently tight confidence interval for detection before and during attack.

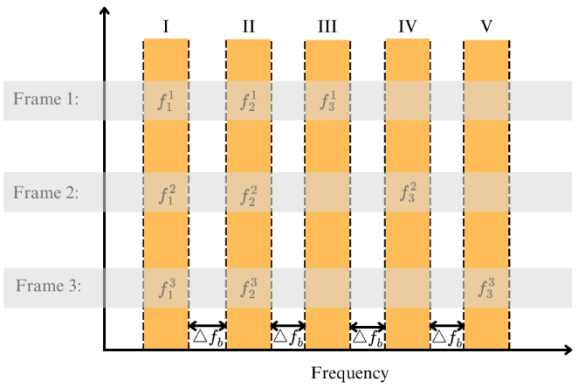
The width of every frequency bin can be very small. In fact, one frequency can be chosen from each $n_m - 1$ bins and assigned to all frames. The $n_m^{th}$ frequency in each frame chosen from a unique frequency bin stands to be the distinguishable frequency. In general, more bins can be considered if desired to have more than one distinguishing frequency per frame.

*Example 3.6.* Suppose the closed loop bandwidth of the system is $f_{BW} = 10\ Hz$. Also, suppose $n_m = 3$. If we choose three frames, i.e. $n_f = 3$, then from (3.36) $n_b \geq 5$. So, we take $n_b = 5$. We choose the frequencies in the range of $1\ Hz$ to $20\ Hz$ and check the frame sizes for the detection of frequencies.

Consider five narrow frequency bins at $1, 6, 10, 15$ and $21\ Hz$ as given in Table 3.1 and depicted in $Fig.$ 3.24. Frequencies $10, 15, 21\ Hz$ are the distinguishing frequencies.

Table 3.1. *Example 3.6.* Frequencies present in each frame.

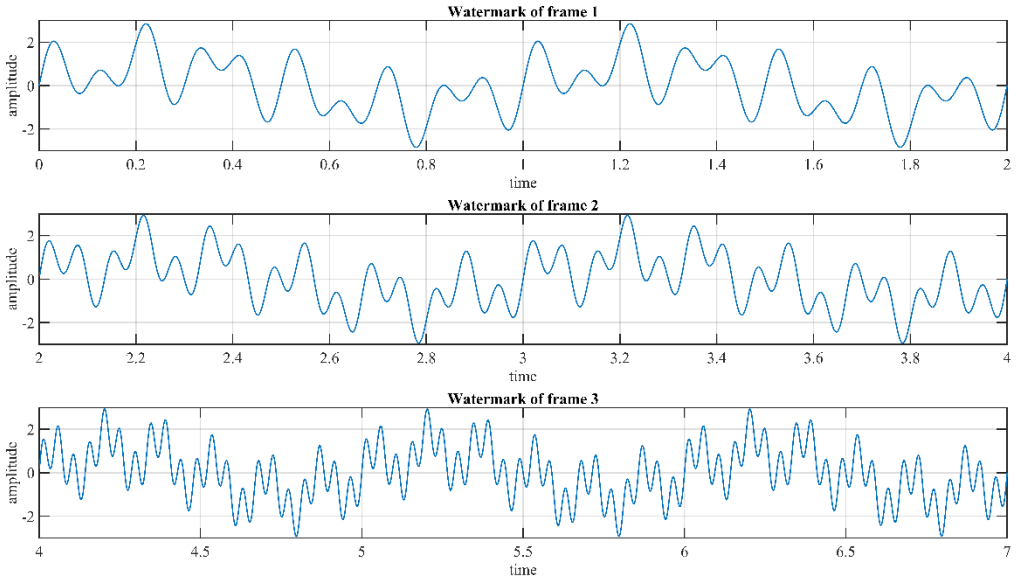
| Frequency ($Hz$) | $f_1^i$ | $f_2^i$ | $f_3^i$ | $T_{f_i}$ | $\Delta f_i$ |
|---|---|---|---|---|---|
| Frame 1, $i = 1$ | 1 | 6 | 10 | $2 * \frac{1}{f_1} = 2$ | $\frac{1}{2}$ |
| Frame 2, $i = 2$ | 1 | 6 | 15 | $2 * \frac{1}{f_1} = 2$ | $\frac{1}{2}$ |
| Frame 3, $i = 3$ | 1 | 6 | 21 | $3 * \frac{1}{f_1} = 3$ | $\frac{1}{3}$ |




$Fig.$ 3.24. Frequency bins for a multi-sine watermarking signal.

$$\Delta f_b = 4 * \max\{\Delta f_1, \Delta f_2, \Delta f_3\} = 2\ Hz$$

Note that $\forall i,j\ (i \neq j), \frac{T_{f_i}}{T_{f_j}} < 2$ and all bins are separated by more than $\Delta f_b = 2\ Hz$. The time domain watermarking signal split in three parts is shown in $Fig.$ 3.25. The periodograms of each frame is shown in $Fig.$ 3.26.




$Fig.$ 3.25. Watermark of each frame with frequencies chosen in Table 3.1.

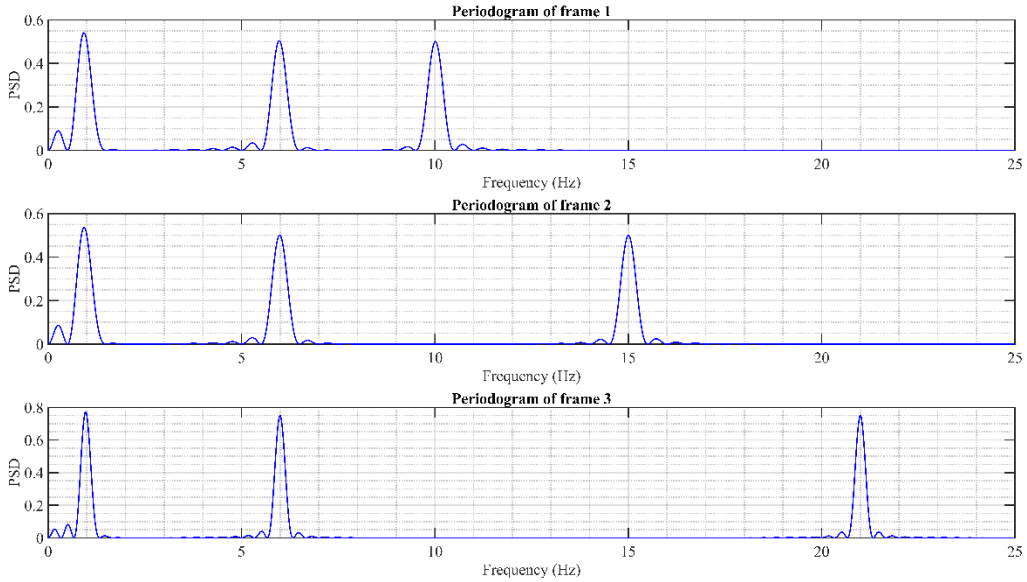



*Fig.* 3.26. Periodogram of each frame with frequencies chosen in Table 3.1.

Therefore, in case of a replay attack, frequencies $10\,Hz$, $15\,Hz$ and $21\,Hz$ will be the distinguishing frequencies for frame 1, 2 and 3, respectively.

*Remark 3.1* The minimum number of frequencies for the multi-sine watermark considered in [44] were $n_f n_m$ as compared to the proposed $n_f + n_m - 1$ in this work. This was concluded as an investigation of the plant output under replay attack was taken into account to design the multi-sine watermark as opposed to [44] where only the normal operation of the plant was considered.

[44] considers distinguishing frequencies of each frame during normal operation, i.e. resolving frequencies in a frame. This thesis also considers distinguishing frequencies in one frame from those in others, i.e. resolving frames from one another. We consider what happens during both normal operation and during attack.

## 3.5 Conclusion

In this chapter, the problem statement was presented to indicate the existing issue that this thesis aims to resolve. A need for the systematic design method of multi-sine watermark was observed. The multi-sine watermark as found in the existing work was analyzed using power spectral analysis under two different cases: system under normal operation and system under replay attack.

The results were utilized to make the necessary modifications in order to make the watermark secure. Finally, a design methodology for the multi-sine watermark was proposed.

In the following chapter, the proposed design method was used to design a multi-sine watermark for a flow control system. This watermark was tested on multiple test cases mirroring real-life scenarios following the same structure of chapter 3. The designed watermark for the case study and its performance analysis are provided in detail in chapter 4.

# Chapter 4

# Case Study: Flow Control System

In this chapter, we design the watermark for a flow control system based on our proposed method presented in the previous chapter. The performance of the multi-sine watermark is then evaluated for multiple cases following the same trend as of chapter 3. The cases are picked to mirror real life scenarios while also giving us a set up for verifying if the watermark enhances the security of the system. We begin by defining our case study model in section 4.1. The proposed design method is followed in section 4.2 to arrive at the desired multi-sine watermark followed by the simulation and analysis of the same in section 4.3. Finally, the conclusion of the analysis is presented in section 4.4.

## 4.1 Plant Model

A ***flow control system*** is considered for the case study. It is a system used to monitor, regulate, and manage the flow of fluids or materials through a network of pipes, channels, or other conduits. The primary function of a flow control system is to ensure that the rate of flow remains within desired parameters, which can be critical in many industrial processes, such as manufacturing, chemical processing, water treatment, and power generation.

A single water tank used in a flow control system is taken as the plant. Figure 4.1 shows the schematic diagram of the tank system.

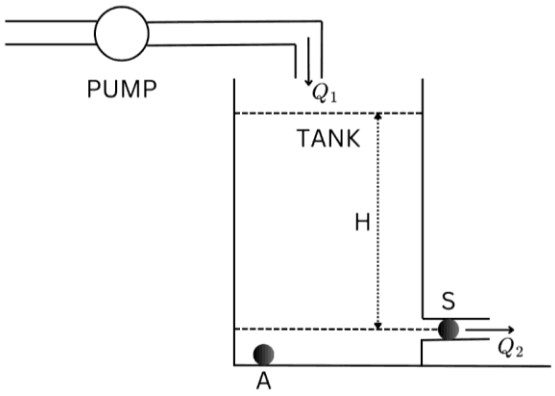



Fig. 4.1. Schematic diagram of single water tank system.

From mass balance equation, we know that

$$A\frac{dH}{dt} = Q_1 - Q_2 \tag{4.1}$$

where $A$ is the tank area, $Q_1$ and $Q_2$ are input and output volume flow, respectively, and $H$ is the water level. The output flow is given by:

$$Q_2 = a_z S(2gH)^{\frac{1}{2}} \tag{4.2}$$

where $a_z$ is the outflow coefficient, $S$ is the cross-sectional area of output pipe and $g$ is acceleration due to gravity. Applying $g = 9.81\, m/s^2$ and $a_z = 0.45,\ S = 5\,X\,10^{-5}\, m^2,\ A = 0.0154\, m^2$ following model AMIRA three tank system DTS200 as given in [26], we get the nonlinear equations of the tank:

$$\begin{cases} \frac{dH}{dt} = -6.49\ \times\ 10^{-3}\sqrt{H} + 64.9Q_1 \\ Q_2 = 0.997\ \times\ 10^{-4}\sqrt{H} \end{cases} \tag{4.3}$$

This tank is part of a flow control system in which $Q_1$ is adjusted to regulate $Q_2$. We use the same operating point values and controller as given in [44] and the model is linearized based on the same operating point.

$$Q_{2_0} = 5.46 \times 10^{-5}\, m^3/s = 54.6\, ml/s \tag{4.4}$$

or

$$H_0 = 0.3\, m. \tag{4.5}$$

Input and output disturbances, $w(t)$ and $v(t)$, respectively, are both assumed to be WSS Gaussian with zero mean and variances

$$\sigma_w^2 = 2 \times 10^{-14}\left(\frac{m^3}{s}\right)^2,\ \sigma_v^2 = 8.25 \times 10^{-15}\left(\frac{m^3}{s}\right)^2 \tag{4.6}$$

respectively. These correspond to standard deviation of $0.14\, ml/s$ and $0.09\, ml/s$. The linearized model around the operating point along with the disturbance is

$$\begin{cases} \frac{dh}{dt} = -5.9 \times 10^{-3}h(t) + 64.9q_1(t) + 64.9w(t) \\ q_2(t) = 9.1 \times 10^{-5}h(t) + v(t) \end{cases} \tag{4.7}$$

Thus, the tank transfer function is

$$G(s) = \frac{1}{169s+1} \tag{4.8}$$

and the controller is a PI controller [44]

$$K(s) = \frac{5.5s+0.1}{s} \tag{4.9}$$

Together, they form the control loop as shown in $Fig.4.2$. In $Fig.4.2$, $r(t)$ is the input, $q_1(t)$ is the input flow deviation from set point, $q_2(t)$ is the output flow (deviation from set point), $m(t)$ is the multi-sine watermarking signal.

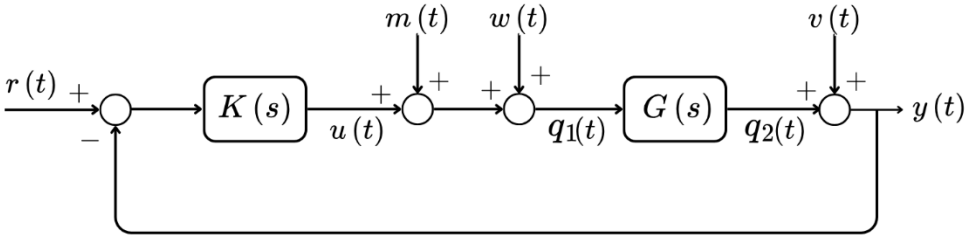



Fig. 4.2. Block diagram of the flow control system.

The step response characteristics of the closed loop system are given in Table 4.1. The bandwidth of the closed-loop system $G_{yr}(s) = \frac{Y(s)}{R(s)}$ is $0.046\ rad/s$.

Table 4.1. Step response characteristics of the flow control system in closed loop.

| **Characteristics** | **Value** |
|---|---|
| Rise Time | $41.48\ s$ |
| Settling Time | $206.46\ s$ |
| Overshoot | $11.85\ \%$ |
| Poles | $-0.0192 \pm j0.0149$ |

From Fig. 4.2, we arrive at the following transfer functions:

$$\frac{Y(s)}{R(s)} = G_{yr}(s) = \frac{5.5s+0.1}{169.2s^2+6.5s+0.1} \tag{4.10}$$

$$G_{yw}(s) = G_{ym}(s) = \frac{s}{169.2s^2+6.5s+0.1} \tag{4.11}$$

$$G_{yv}(s) = \frac{169.2s^2+s}{169.2s^2+6.5s+0.1} \tag{4.12}$$

$$G_{ur}(s) = \frac{929.5s^2+22.42s+0.1}{169.2s^2+6.5s+0.1} \tag{4.13}$$

## 4.2 Proposed Design of Watermark

Watermarking signal over a frame is given by:

$$m(t) = c_0 \sum_{i=1}^{n_m} A_i \sin(\omega_i t + \phi_i) \tag{4.14}$$

where $t$ is measured from the start of time, $n_m$ is the number of unique frequencies within a frame and $T_f$ is the frame length. Furthermore, $A_i$, $\omega_i$ and $\phi_i$ represent the amplitude, frequency and phase of the sine components, respectively. For each frame, the frequencies in $Hz$ and periods are denoted by $f_i = \frac{\omega_i}{2\pi}$ and $T_i = \frac{1}{f_i}$, respectively (here, we have chosen $c(s)$ in (2.6) as $c(s) = c_0$).

Following (2.6) from chapter 2 with $c(s) = c_0$, we have $n_m \geq n$ and the order of the closed-loop system is $n = 2$. We choose $n_m = 2$ and include two frequencies in each frame. To explore all possible attack and detection combinations, the number of frames was kept three ($n_f = 3$) similar to [44]. Therefore,

$$m_1(t) = c_0^1 \times (A_1 \sin(\omega_1 t + \phi_1) + A_2 \sin(\omega_2 t + \phi_2)) \quad \text{(frame 1)} \tag{4.15}$$

$$m_2(t) = c_0^2 \times (A_3 \sin(\omega_3 t + \phi_3) + A_4 \sin(\omega_4 t + \phi_4)) \quad \text{(frame 2)} \tag{4.16}$$

$$m_3(t) = c_0^3 \times (A_5 \sin(\omega_5 t + \phi_5) + A_6 \sin(\omega_6 t + \phi_6)) \quad \text{(frame 3)} \tag{4.17}$$

Here, for each watermark, $t$ is with respect to the start of frame. Based on the discussion in section 3.4, we consider $n_b = 4$. The bandwidth of the closed-loop system is $0.0446\ rad/s$. To have short frame size, we choose watermark frequencies from around bandwidth and higher. To avoid high

frequency noise and modeling uncertainty, we choose frequencies less than 10 times of bandwidth. Thus, we choose the frequencies $(rad/s)$ to be $0.06, 0.18, 0.30, 0.42$. The assignment of frequencies, the corresponding integers $n_i$, amplitude $A_i$ and phase $\phi_i$ are given in Table 4.2. Note that the first frequency $(0.06\, rad/s)$ is common and the other three serve as the distinguishing frequencies.

Table 4.2. Values proposed for the watermark.

| | Frame 1 | | Frame 2 | | Frame 3 | |
|---|---|---|---|---|---|---|
| Variable | sinusoid 1 | sinusoid 2 | sinusoid 3 | sinusoid 4 | sinusoid 5 | sinusoid 6 |
| $n_i$ | 1 | 3 | 1 | 5 | 1 | 7 |
| $\omega_i$ | 0.06 | 0.18 | 0.06 | 0.30 | 0.06 | 0.42 |
| $A_i$ | 371.14 | 1062.50 | 123.71 | 588.47 | 61.86 | 411.59 |
| $\phi_i$ | 2.4879 | −0.2141 | 2.4879 | −0.1282 | 2.4879 | −0.0915 |

For frame 1, $T_{comb} = \frac{2\pi}{0.06}$ and we choose $T_{f_1} = 3T_{comb} = 100\pi = 314s$.

For frame 2, $T_{comb} = \frac{2\pi}{0.06}$ and we choose $T_{f_2} = 2T_{comb} = \frac{200}{3}\pi = 209s$.

For frame 3, $T_{comb} = \frac{2\pi}{0.06}$ and we choose $T_{f_3} = 2T_{comb} = \frac{200}{3}\pi = 209s$.

Therefore, frame 1 begins at $t = 0s$ and ends at $t = 314s$, frame 2 begins at $t = 314s$ and ends at $t = 523s$ and frame 3 begins at $t = 523s$ and ends at $t = 732s$. Resolution,

$$\Delta f = \max_{\mathrm{i}} \Delta f_i = \max_{\mathrm{i}} \frac{1}{T_{f_i}} = 0.0048\ Hz$$

and

$$4\Delta f = 0.0192\ Hz$$

or

$$4\Delta\omega = 0.12\ rad/s$$

Therefore, the frequency bins are appropriately placed. Now, suppose that the maximum allowable fluctuation in water level due to watermarking is $2\ cm$, or $6.7\%$ of the operating point. Then the maximum output flow fluctuation is $\delta_m = 1.8$ ml/s, which corresponds to 3.3% of the operating point value ($54.6\ ml/s$). During every frame, the impact of watermarking is

$$c_0 \sum_{i=1}^{n_m} A_i \left| G_{ym}(j\omega_i) \right| \sin(\omega_i t + \phi_i + \measuredangle G_{ym}(j\omega_i))$$

An upper bound for the above is

$$c_0 \sum_{i=1}^{n_m} A_i \left| G_{ym}(j\omega_i) \right|$$

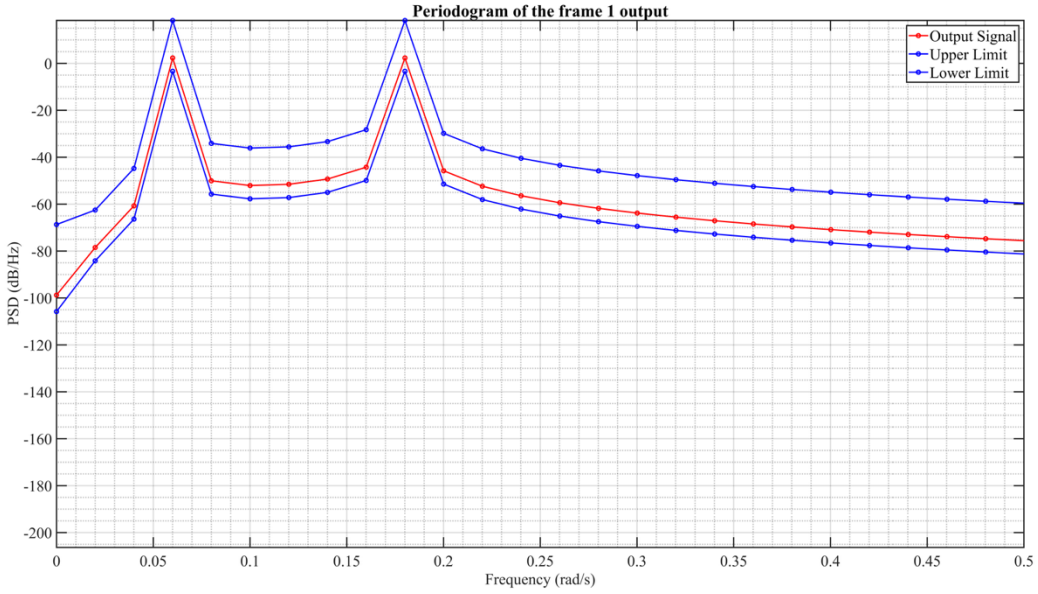
We select $c_0^i$ such that

$$c_0^1 \left( A_1 \left| G_{ym}(j\omega_1) \right| + A_2 \left| G_{ym}(j\omega_2) \right| \right) = c_0^1 \times 69.4596 \leq \delta_m \qquad \text{(frame 1)} \qquad (4.18)$$

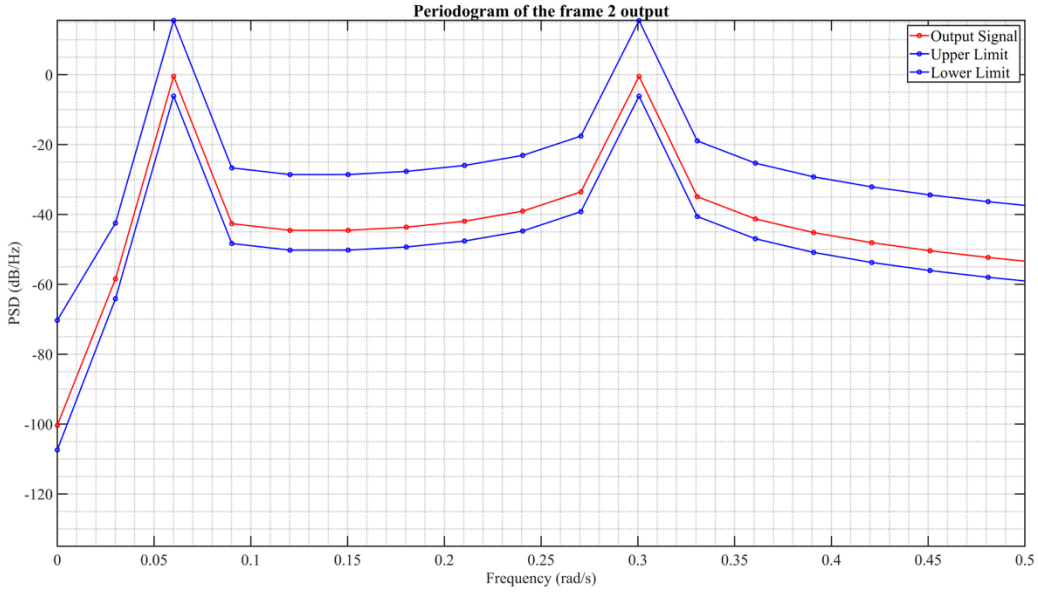
$$c_0^2 \left( A_3 \left| G_{ym}(j\omega_3) \right| + A_4 \left| G_{ym}(j\omega_4) \right| \right) = c_0^2 \times 23.1525 \leq \delta_m \qquad \text{(frame 2)} \qquad (4.19)$$

$$c_0^3 \left( A_5 \left| G_{ym}(j\omega_5) \right| + A_6 \left| G_{ym}(j\omega_6) \right| \right) = c_0^3 \times 11.5767 \leq \delta_m \qquad \text{(frame 3)} \qquad (4.20)$$

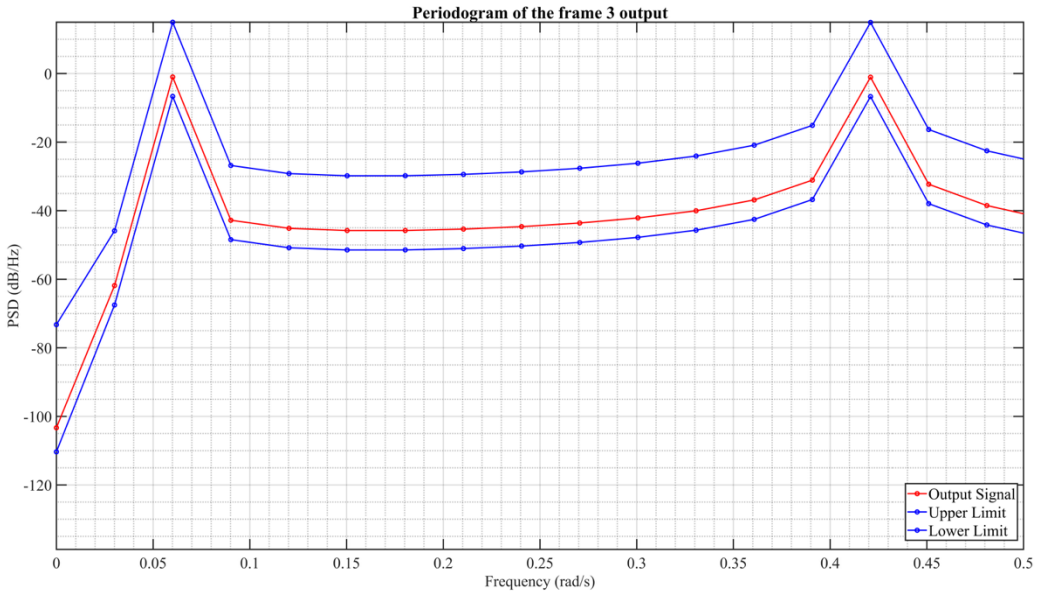
Following section 3.2, $c_0^1 \leq 0.026$, $c_0^2 \leq 0.078$ and $c_0^3 \leq 0.155$.The values of $c_0^i$ are chosen so the peaks of periodogram are in the $95\%$ confidence bound region and the lower bound exceeds the upper bound by at least $+3dB$ at the expected frequencies. When the system of $Fig.4.2$ in steady state is under safe operation, the frame-wise detection using $95\%$ confidence bound periodogram with $T_s = 0.01s$ as shown in $Fig.4.3$ is used to tune the values $c_0^1 = 0.003$, $c_0^2 = 0.008$ and $c_0^3 = 0.015$ to maintain a difference of at least $+8\ dB$ at the expected frequencies. For example, in $Fig.4.3(a)$, the lower bounds at frequencies $0.06\,rad/s$ and $0.18\,rad/s$ are $-3\ dB/(rad/s)$ whereas at other frequencies, the upper bounds are below $-28\ dB/(rad/s)$.


Periodogram of the frame 1 output
Output Signal
Upper Limit
Lower Limit
PSD (dB/Hz)
Frequency (rad/s)

$(a)$


Periodogram of the frame 2 output
Output Signal
Upper Limit
Lower Limit
PSD (dB/Hz)
Frequency (rad/s)

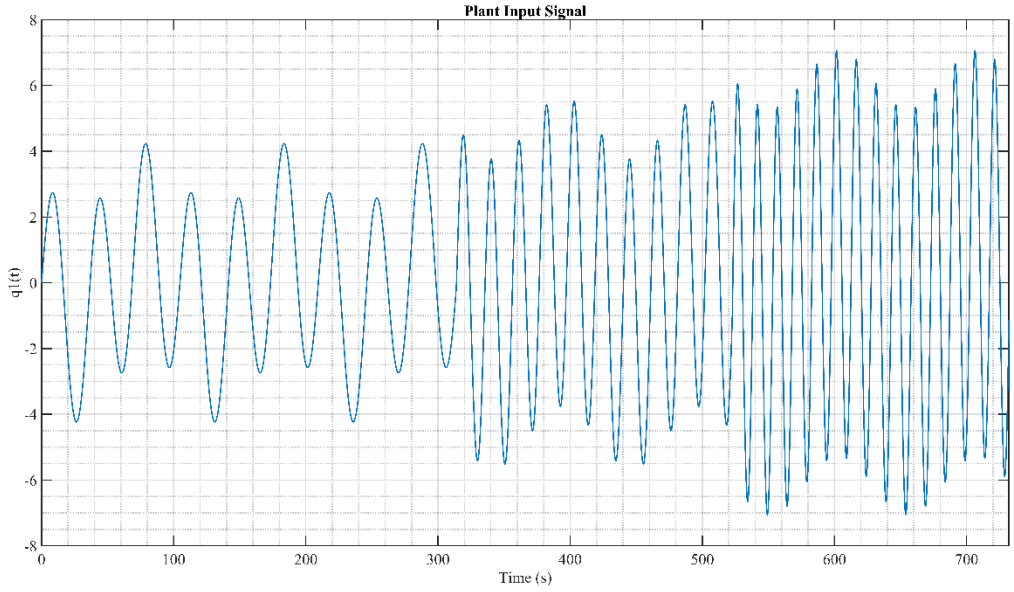
$(b)$




$(c)$

$Fig.$ 4.3. Periodogram of the output simulated under safe conditions for $(a)$ frame 1, $(b)$ frame 2, and $(c)$ frame 3, respectively.

The resulting control input to the plant for the case study is shown in $Fig.$ 4.4.




$Fig$ 4.4. Plant input with multi-sine watermarking signal over three frames.

In the following section, a detailed analysis of the case study simulated along with the designed multi-sine watermark is presented. The detection of the watermark before and during a replay attack are studied in different scenarios following the format discussed in chapter 3.

## 4.3 Simulation Results

The system discussed in the previous section is simulated using MATLAB. The system described in $Fig.\,4.2$ is simulated and is divided into cases as discussed in chapter 3. In section 4.2.1, the system operating under safe/normal conditions is covered and in section 4.2.2, the system under a replay attack is analyzed. The 95% confidence bound periodogram is used for detecting the frequencies present in the signal.

### 4.3.1 Before Replay Attack

In this section, the analysis of the system with multi-sine watermarking under normal conditions is reiterated. The closed loop system is considered to be in in steady state. When this system is under safe operation, i.e. it is not under a replay attack, the frame-wise detection using 95% confidence bound periodogram is as was shown in $Fig.\,4.3$.

We can observe that at the expected frequencies (i.e. at $0.06\ Hz$ and $0.18\ Hz$ for frame 1, $0.06\ Hz$ and $0.30\ Hz$ for frame 2 and $0.06\ Hz$ and $0.42\ Hz$ for frame 3), the lower bound of the 95% confidence bound periodogram is above the upper bound at other frequencies. Hence, the expected frequencies in each frame are detected according to the 95% confidence bound criteria for each frame, respectively. Therefore, the detection mechanism returns the result that the system is free from a replay attack and operating under safe conditions with a false alarm rate of 5%.

### 4.3.2 During Replay Attack

In this section, the closed loop system with watermark $m(t)$, disturbance signals $w(t)$ and $v(t)$ present is considered. It is assumed that an attacker records and replays a segment of the output, hence, the system with multi-sine watermarking is under a replay attack. The frame-wise detection using 95% confidence bound periodogram varies according to the nature of the replayed signal. Therefore, as discussed in section 3.3 of this thesis, two cases are considered, namely, $\delta \ll T_{f_i}$ dealt with in sub-section 4.2.2.1 and $\delta \gg T_{f_i}$ analyzed in sub-section 4.2.2.2. It should be noted

that the former case is expected in process control but not the latter. In this section, we also present the second case to illustrate the discussion in chapter 3.

### *4.3.2.1 Short Repeated Segments*

*Analysis* 1: For case (1) when $\delta \ll T_{f_i}$, we analyze frame 1 under replay attack. The replayed data is $1/10th$ of frame 3 recorded and repeated in frame 1. The attack commences and lasts for $314s$ during the entire frame. Figure 4.5 shows the replayed signal in time domain.

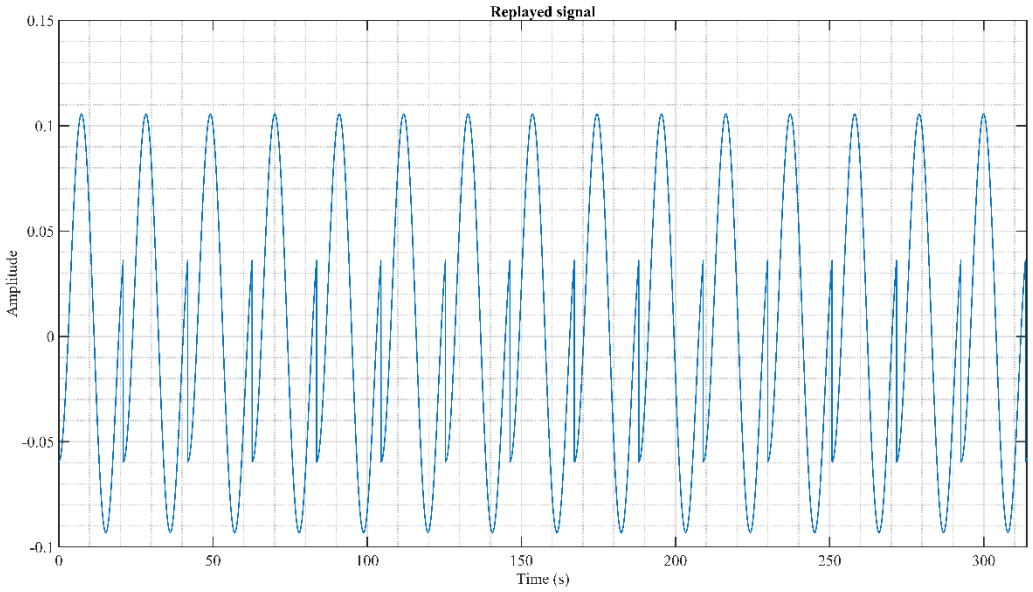



*Fig.* 4.5. *Analysis 1:* Time domain representation of $1/10th$ of frame 3 repeated and replayed in frame 1.

The output frequencies detection using 95% confidence bound periodogram is shown in Fig. 4.6.

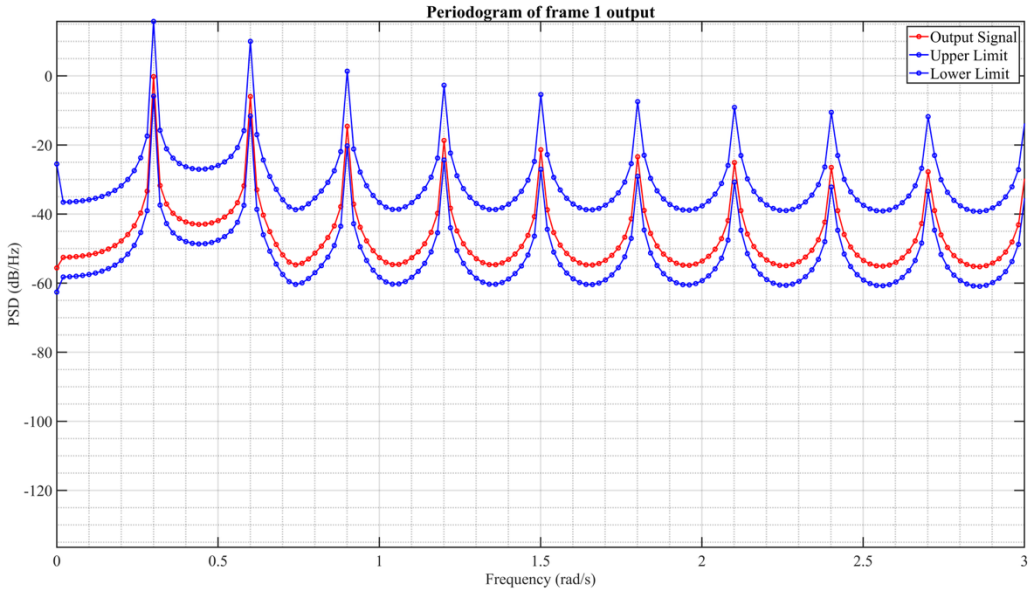



*Fig*. 4.6. *Analysis 1:* Periodogram of the frame 1 output under replay attack (replay of frame 3).

It is evident from *Fig*. 4.6 that when the system is under attack, for this case, frequency $\frac{2\pi}{\delta} = \frac{2\pi}{20.9} \approx 0.3\ rad/s$ and its multiple harmonic frequencies are detected since the repeated signal introduces these in the frequency spectrum. Also, we note a drop of $\sim 5dB$ in the PSD of the base harmonic compared to the original response in frame 3, as is expected in line with the discussion of section 3.3.1. Therefore, the presence of unexpected frequencies concludes the presence of a replay attack. It is noteworthy that the amplitude of the harmonics decreases as they diverge from the base frequency and eventually the PSD at the harmonic frequencies goes undetected by 95% confidence bound periodogram.

*Analysis* 2: For case (1) when $\delta \ll T_{f_i}$, we analyze frame 2 under replay attack. The replayed data has been recorded from $1/10th$ of frame 1 and repeated for the duration of entire frame 2. Figure 4.7 shows the time domain representation of the replayed signal.

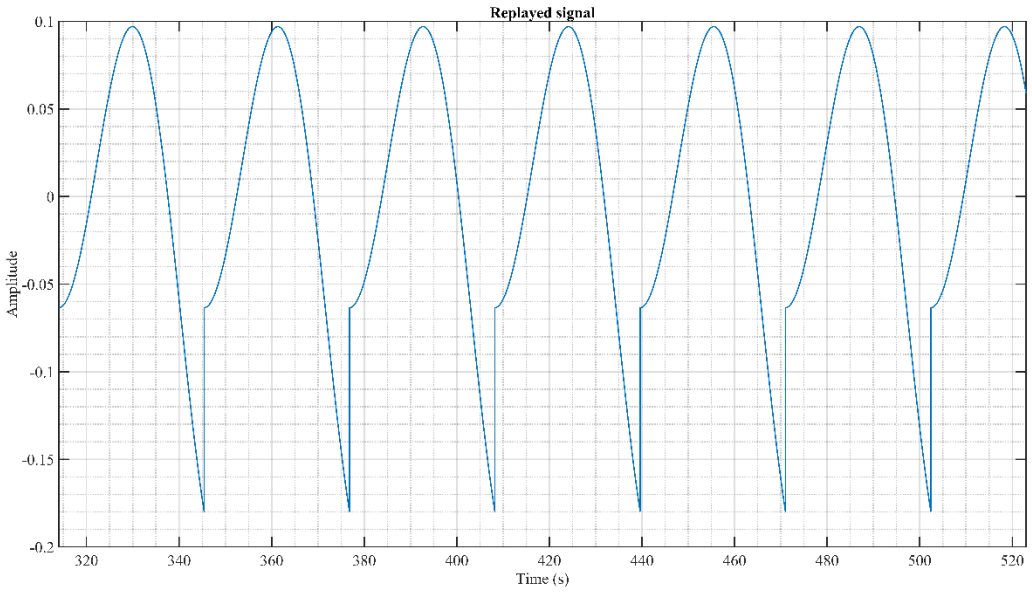



*Fig*. 4.7. *Analysis 2:* Time domain representation of $1/10th$ of frame 1 repeated and replayed in frame 2.

The 95% confidence bound periodogram of the output signal is shown in *Fig*. 4.8.

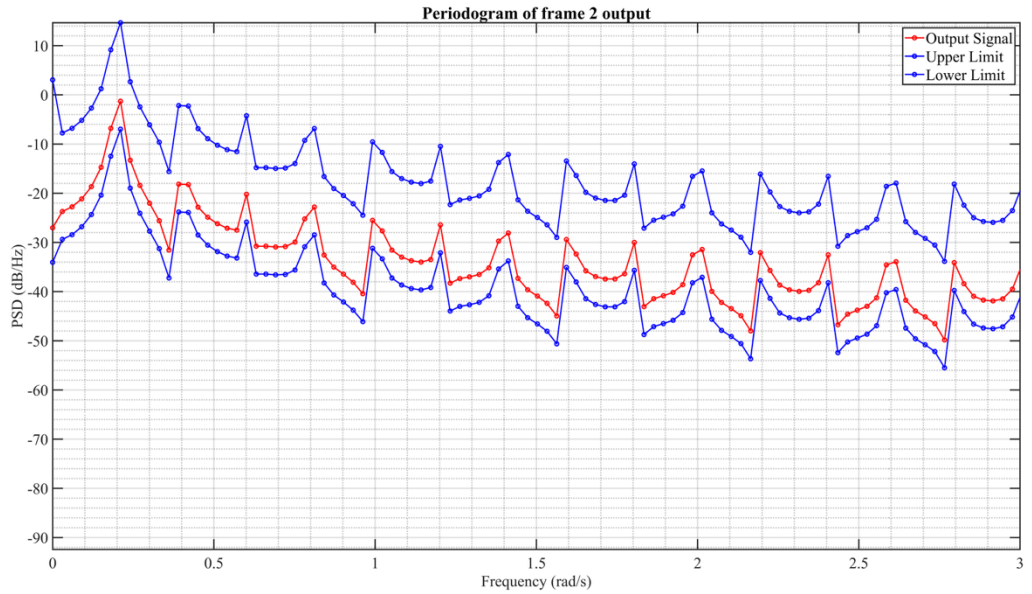



*Fig*. 4.8. *Analysis 2:* Periodogram of the frame 2 output under replay attack (replay of frame 1).

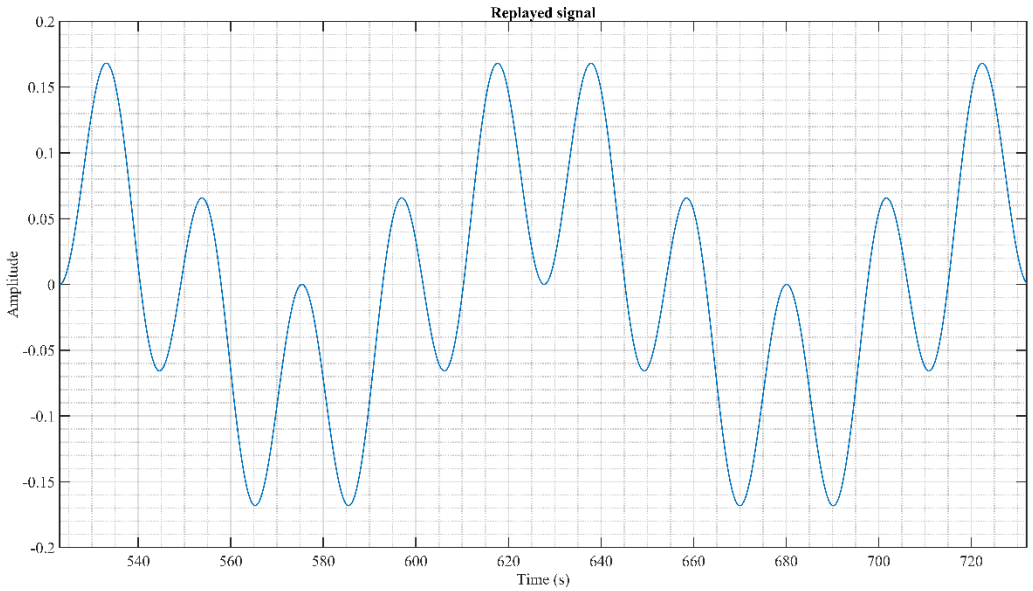
Similar to the result of Analysis 1, it can be observed that frequency $\frac{2\pi}{\delta} = \frac{2\pi}{31.4} \approx 0.2\ rad/s$ and its multiple harmonic frequencies are present in the output signal based on the peaks observed in the periodogram in *Fig*. 4.8. However, in this case, they remain undetected in the 95% confidence

bounds. There is a reduction of $\sim 16\ dB$ in the PSD of the base harmonic as discussed in section 3.3.1. Therefore, there is no successful detection of any frequencies including frame 2 frequencies. The absence of the expected unique frequency of frame 2 at $\omega_4 = 0.30\ rad/s$ confirms that the system is undergoing a replay attack.

### *4.3.2.2 Long Replayed Segments*

For case $(2)$ when $\delta \gg T_{f_i}$, we analyze the sub cases $(2A)$, $(2B)$ and $(2C)$ as discussed in section 3.3.

*Analysis* 3: Case $(2A)$: Consider frame 3 under attack with replayed data that was recorded from frame 2 with $\alpha = 1$. The attack lasts for the whole duration of the frame. The recorded signal is assumed to be as shown in $Fig.\ 4.9$.




$Fig.\ 4.9$. *Analysis 3:* Time domain representation of frame 2 output replayed in frame 3.

The 95% confidence bound periodogram of the output signal is shown in $Fig.\ 4.10$.

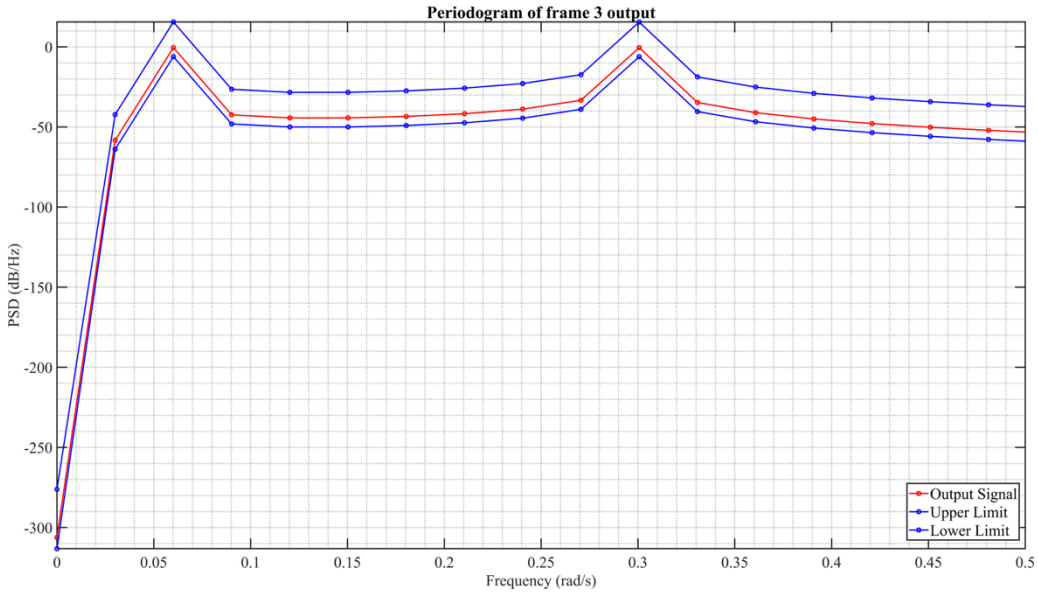



*Fig*. 4.10. *Analysis 3:* Periodogram of the frame 3 output under replay attack (replay of frame 2).

Since the whole frame is under a replay attack and the replayed data was recorded from frame 2, which was a frame of equal size, we observe peak at the frame 2 unique frequency ($\omega_5 = 0.30$) detectable in the 95% confidence bound periodogram. Therefore, the detection of unexpected frequency can be used to confirm the occurrence of a replay attack.

*Analysis* 4 Case $(2A)$: Consider frame 2 under attack with replayed data that was recorded from frame 1 ($92s$ to $301s$). The attack lasts for the entire duration of the frame. With $\alpha = 0.67$, the replayed signal is assumed to be as shown in *Fig*. 4.11.

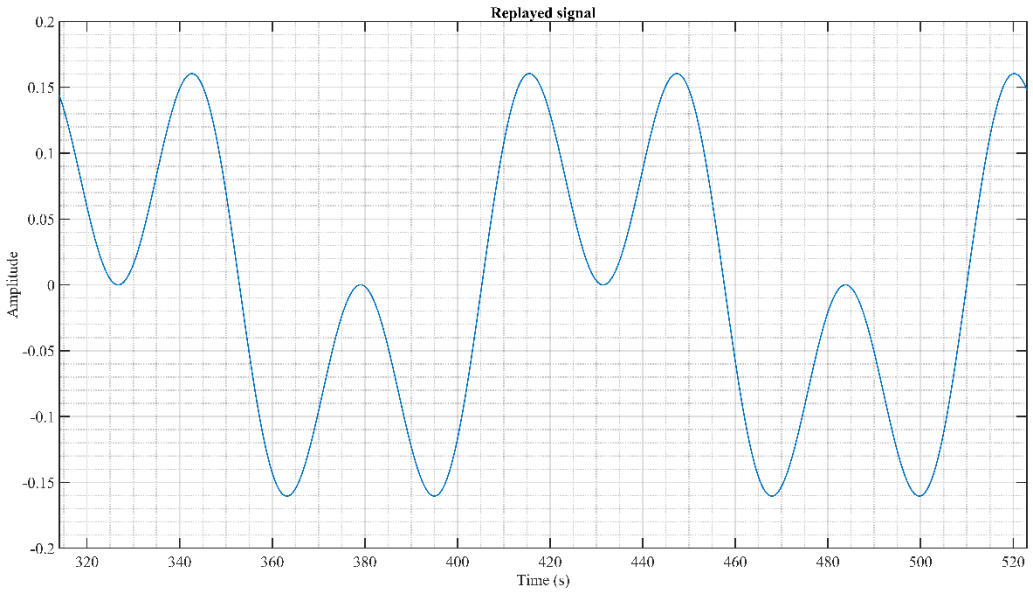



*Fig*. 4.11. *Analysis 4:* Time domain representation of frame 1 output replayed in frame 2.

The 95% confidence bound periodogram of the output signal is shown in *Fig*. 4.12.

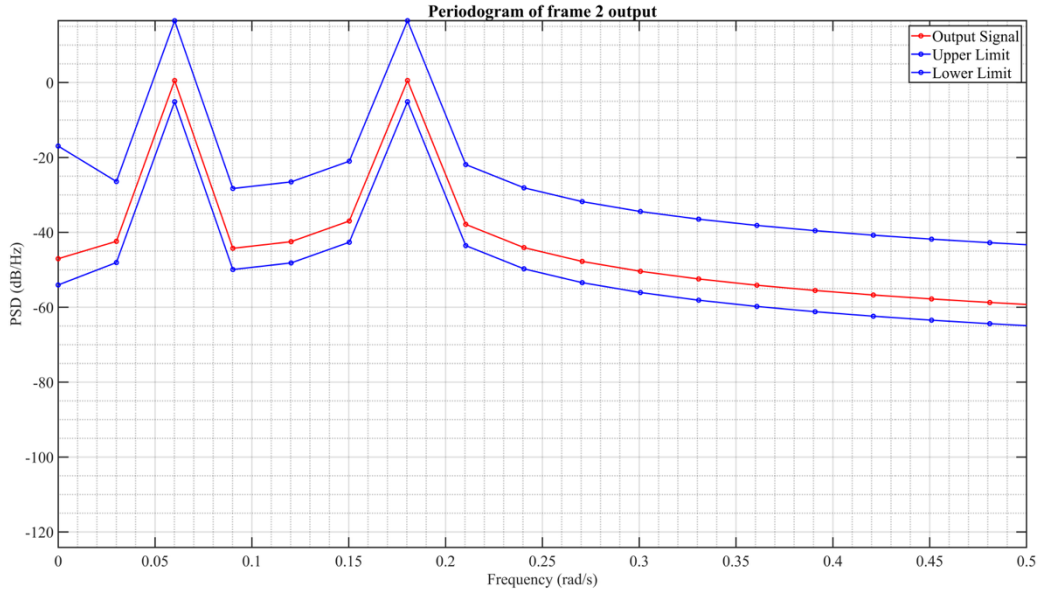
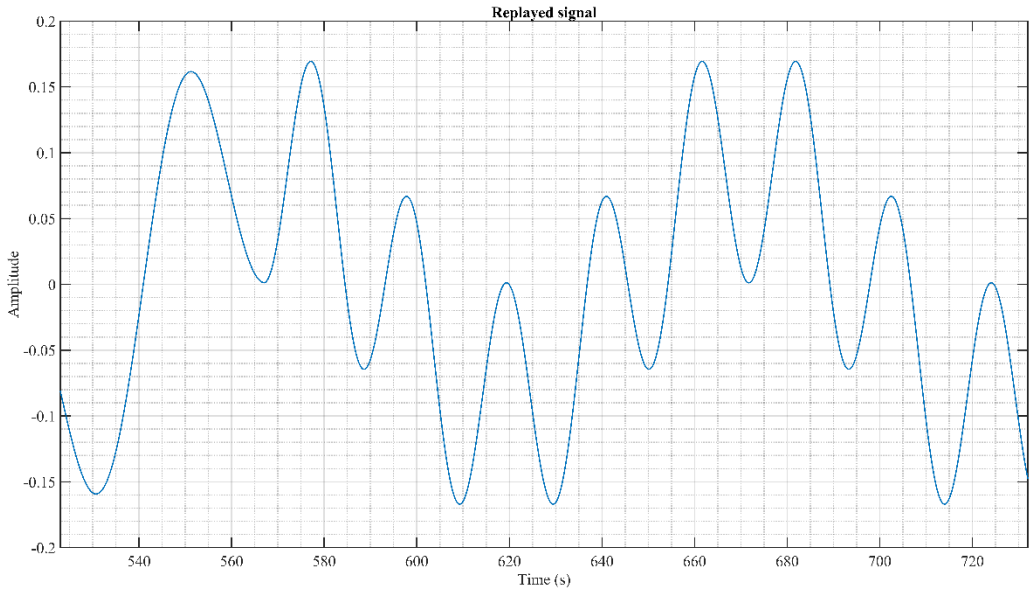



*Fig*. 4.12. *Analysis 4:* Periodogram of the frame 2 output under replay attack (replay of frame 1).

Since with $\alpha = 0.67$, the whole frame is under a replay attack and the replayed data was recorded from frame 1, we observe a peak around the differentiating frequency of frame 1 ($\omega_2 = 0.18$) in the periodogram. Despite the amplitude being dampened with a loss in resolution from $0.02\ rad/s$ to $0.03\ Hz$, frame 1 unique frequency at $0.18\ rad/s$ is successfully detected in the 95%

confidence bound periodogram. Therefore, the detection of unexpected frequency can be used to confirm the occurrence of a replay attack.

*Analysis* 5: Case $(2B)$: Consider frame 3 under attack with replayed data that was recorded from frames 1 and 2 $(270s$ to $479s)$ with $\alpha_1 = 0.14$ and $\beta = 0.21$. The attack lasts for the whole duration of the frame. The recorded signal is assumed to be as shown in $Fig.4.13$.




$Fig.4.13$. *Analysis 5:* Time domain representation of output of frames 1 and 2 replayed in frame 3.

The 95% confidence bound periodogram of the output signal is shown in $Fig.4.14$.

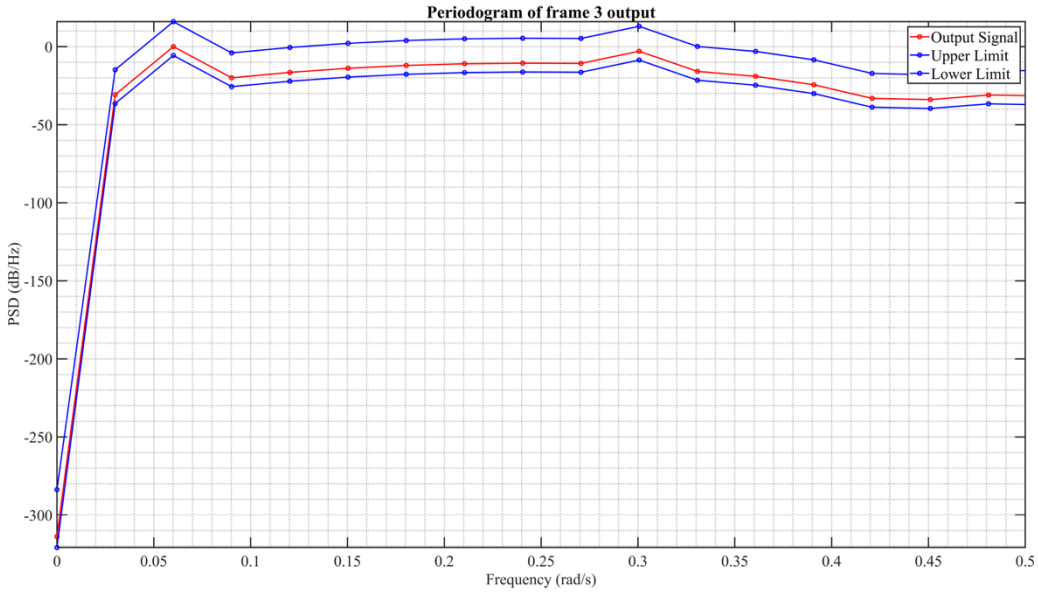



*Fig*. 4.14. *Analysis 5:* Periodogram of the frame 3 output under replay attack (replay of frames 1 and 2).

As the data recorded from two different frames was replayed over frame 3, there were more than two frequencies present in the frame. Although the differentiating frequency of frame 2 ($0.30\,rad/s$) shows a peak, the presence of multiple frequencies stretches and dampens the amplitude in the 95% confidence bound periodogram. Anyways, the expected frequencies are not found and confirm the presence of a replay attack.

*Analysis* 6: Case $(2B)$: Consider frame 3 under attack with replayed data that was recorded from frames 1 and 2 ($210s$ to $419s$) with $\alpha_1 = 0.33$ and $\beta \sim 0.5$. The attack lasts for the whole duration of the frame. The recorded signal is assumed to be as shown in *Fig*. 4.15.

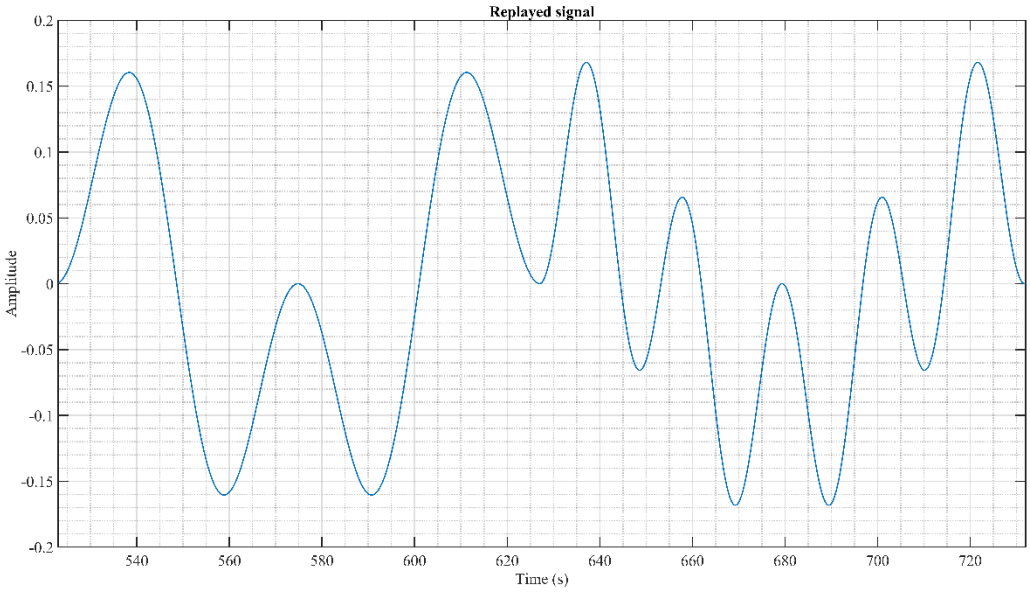



*Fig*. 4.15. *Analysis 6:* Time domain representation of output of frames 1 and 2 replayed in frame 3.

The 95% confidence bound periodogram of the output signal is shown in *Fig*. 4.16.

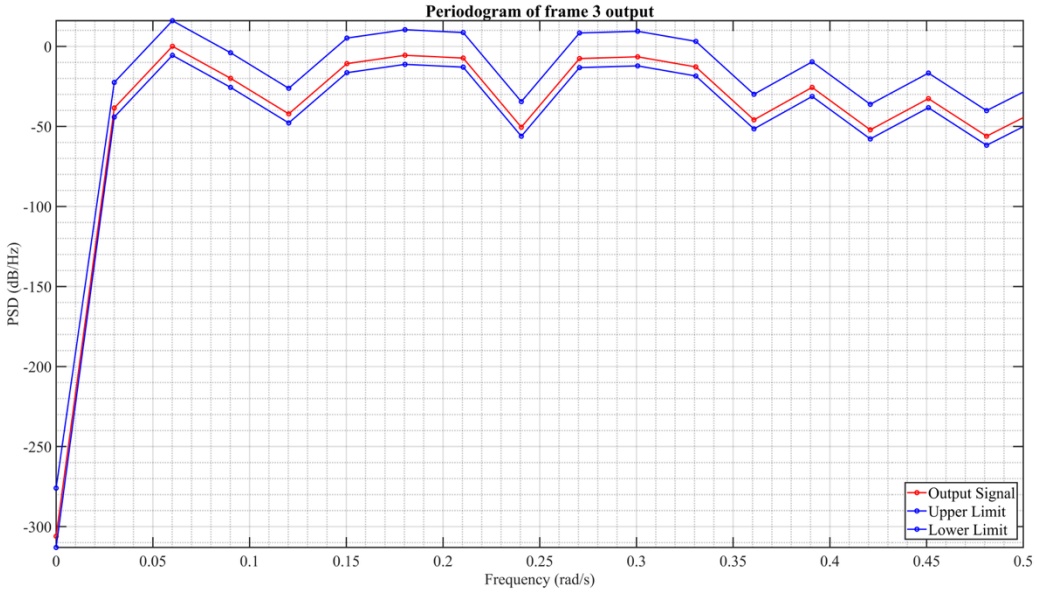



*Fig*. 4.16. *Analysis 6:* Periodogram of the frame 3 output under replay attack (replay of frames 1 and 2).

In this case, the PSD is stretched and dampened even more than in *Analysis* 5. A dome is observed over the frequencies of our concern. Therefore, a similar conclusion as of *Analysis* 5 is drawn since

the data recorded from two different frames was replayed over frame 3. The expected frequencies are not found and confirm the presence of a replay attack.

## 4.4 Conclusion

Replay attack is one of the easiest forms of cyber-attacks on control systems as the intruder needs little information about the system dynamics. This attack is successfully detected by a multi-sine watermarking signal that is designed based on the procedure using power spectral analysis as proposed in this thesis. This watermarking signal that is used as the authentication signal is verified using a periodogram to detect the frequencies present in the signal. The 95% confidence bound periodogram successfully authenticates the signal aimed at detecting a replay attack at early stages. Various cases mirroring the real-life scenarios wherein a replay attack may be carried out starting from different points of time with altering durations within a single frame and across multiple frames were considered and the detection mechanism succeeded in detecting the replay attack.

# Chapter 5

# Conclusion and Future Research

## 5.1 Conclusion

This thesis presented an algorithm for the design of a multi-sine watermarking signal used for the detection of replay attacks. To be precise, the parameters of the multi-sine watermarking signal, namely number of frames, number of watermark sinusoids in each frame, watermark frequencies for each frame and the scaling factor $c_0$ were designed using power spectral analysis. The confidence bounds of periodogram were used to investigate the output for the presence of the watermark frequencies during normal operation and during a replay attack and to adjust the false alarm rate.

The periodogram was used to detect the watermark frequencies in the appropriate frames under normal operation. However, during a replay attack, the analysis of output was divided into two different cases. In case (1), the replayed segment consisted of a small portion of the recorded signal replayed repeatedly. The periodogram detected unexpected frequencies dependent on the length of the recorded portion in the output, thereby indicating a replay attack. In case (2), a long portion of the recorded signal was replayed. The periodogram of the frame confirmed the absence of the expected frequencies and the presence of unexpected frequencies with a reduced resolution in some cases. Therefore, in either case, the presence of unexpected frequencies or the absence of expected frequencies confirmed a replay attack in progress.

Through detailed case study simulations and analysis, the design process of the multi-sine watermarking signal was assessed, showing its ability to withstand replay attacks.

## 5.2 Future Work

While this research has demonstrated the effectiveness of the multi-sine watermarking technique for replay attack detection, several areas remain open for future exploration.

1. Measurement and reduction of detection time using the proposed watermark could also be explored to improve efficiency.

2. An extension of the multi-sine watermarking technique would be valuable to multi-input multi-output (MIMO) systems.
3. Development of adaptive watermarking schemes that adjust the multi-sine parameters based on real-time feedback from the system. This dynamic adaptability could significantly enhance the technique's ability to detect more sophisticated and evolving attacks.
4. Further exploration into the integration of machine learning or artificial intelligence algorithms could further enhance the technique's detection accuracy and allow it to learn from attack patterns over time.
5. Finally, experimental validation of the technique in real-world environments would provide valuable insights into its performance and effectiveness, helping to ensure that it can meet the demands of practical, large-scale implementations.

# References


[1] A. A. Cardenas, S. Amin and S, Sastry, “Research Challenges for the Security of Control Systems,” *3rd USENIX Workshop on Hot Topics in Security, 2008*.

[2] M. Asiri, N. Saxena, R. Gjomemo and P. Burnap, “Understanding Indicators of Compromise against Cyber-attacks in Industrial Control Systems: A Security Perspective,” *ACM Transactions on Cyber-Physical Systems*, vol. 7, no. 2, pp. 1-33, 2023, doi: 10.1145/3587255.

[3] S. M. Dibaji, M. Pirani, D. B. Flamholz, A. M. Annaswamy, K. H. Johansson and A. Chakrabortty, "A Systems and Control Perspective of CPS Security," *Annual Reviews in Control*, vol. 47, pp. 394-411, 2019, doi: 10.1016/j.arcontrol.2019.04.011.

[4] J. Li, Z. Wang, Y. Shen and L. Xie, “Attack Detection for Cyber-Physical Systems: A Zonotopic Approach,” *IEEE Transactions on Automatic Control*, vol. 68, no. 11, pp. 6828-6835, 2023, doi: 10.1109/TAC.2023.3240383.

[5] S. Karnouskos, "Stuxnet worm impact on industrial cyber-physical system security," *IECON 2011 - 37th Annual Conference of the IEEE Industrial Electronics Society,* Melbourne, VIC, Australia, pp. 4490-4494, 2011, doi: 10.1109/IECON.2011.6120048.

[6] M. A. Ferrag, I. Kantzavelou, L. Maglaras, and H. Janicke, *Hybrid Threats, Cyberterrorism and Cyberwarfare,* 1st ed., CRC Press, 2024, doi: 10.1201/9781003314721.

[7] Y. Mekdad, G. Bernieri, M. Conti, and A. E. Fergougui, “A threat model method for ICS malware: the TRISIS case,” *Proceedings of the 18th ACM International Conference on Computing Frontiers (CF '21)*, Association for Computing Machinery, New York, NY, USA, pp. 221–228, 2021, doi: 10.1145/3457388.3458868.

[8] A. Amin, A. A. Cardenas and S. Sastry, “Safe and secure networked control systems under denial-of-service attacks,” *Proceedings of the Hybrid Systems: Computation and Control, LNCS 5496*, Springer-Verlag, pp. 31-45, 2009, doi: 10.1007/978-3-642-00602-9_3

[9] Y. Mo and B. Sinopoli, “False data injection attacks in control systems,” *Proceedings of the 1st Workshop Secure Control Systems*, Stockholm, Sweden, 2010.

[10] D. Bhamare, M. Zolanvari, A. Erbad, R. Jain, K. Khan and N. Meskin, “Cybersecurity for industrial control systems: A Survey,” *Computers & Security*, vol. 89, 2020, doi: 10.1016/j.cose.2019.101677.

[11] T. Alladi, V. Chamola and S. Zeadally, “Industrial Control Systems: Cyberattack trends and countermeasures,” *Computer Communications*, vol. 155, pp. 1-8, 2020, doi: 10.1016/j.comcom.2020.03.007.

[12] T. Miller, A. Staves, S. Maesschalck, M. Sturdee and B. Green, “Looking back to look forward: Lessons learnt from cyber-attacks on Industrial Control Systems” *International Journal of Critical Infrastructure and Protection*, vol. 35, 2021, doi: 10.1016/j.ijcip.2021.100464.

[13] Z. Drias, A. Serhrouchni and O. Vogel, "Analysis of cyber security for industrial control systems," 2015 International Conference on Cyber Security of Smart Cities, *Industrial Control System and Communications (SSIC)*, Shanghai, China, pp. 1-8, 2015, doi: 10.1109/SSIC.2015.7245330.

[14] X. Fan, K. Fan, Y. Wang and R. Zhou, "Overview of cyber-security of industrial control system," *International Conference on Cyber Security of Smart Cities, Industrial Control System and Communications (SSIC)*, Shanghai, China, pp. 1-7, 2015, doi: 10.1109/SSIC.2015.7245324.

[15] S. McLaughlin, C. Konstantinou, X. Wang, L. Davi, A. R. Sadeghi, M. Maniatakos and R. Karri, "The Cybersecurity Landscape in Industrial Control Systems," in *Proceedings of the IEEE*, vol. 104, no. 5, pp. 1039-1057, 2016, doi: 10.1109/JPROC.2015.2512235.

[16] Y. Mo, S. Weerakkody and B. Sinopoli, “Physical authentication of control systems: designing watermarked control inputs to detect counterfeit sensor outputs, ” *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 93-109, 2015, doi: 10.1109/MCS.2014.2364724.

[17] A. Khazraei, H. Kebriaei and F. R. Salmasi, “A New Watermarking Approach for Replay Attack Detection in LQG Systems,” *56th IEEE Annual Conference on Decision and Control*, 2017, doi: 10.1109/CDC.2017.8264421.

[18] C. Trapiello and V. Puig, “Set-based replay attack detection in closed-loop systems using a plug & play watermarking approach,” *4th Conference on Control and Fault Tolerant Systems*, 2019, doi: 10.1109/SYSTOL.2019.8864790.

[19] S. Weerakkody, Y. Mo and B. Sinopoli, “Detecting Integrity Attacks on Control Systems using Robust Physical Watermarking,” *53rd IEEE Conference on Decision and Control*, 2014, doi: 10.1109/CDC.2014.7039974.

[20] R. Romagnoli. S. Weerakkody and B. Sinopoli, “A Model Inversion Based Watermark for Replay Attack Detection with Output Tracking,” *American Control Conference,* 2019, doi: 10.23919/ACC.2019.8814483.

[21] H. Liu, Y. Mo, J. Yan, L. Xie and K. H. Johansson, “An on-line approach to physical watermark design,” *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3895-3902, 2020, doi: 10.1109/TAC.2020.2971994.

[22] C. M. Ahmed, V. R. Palleti and V. K. Mishra, “A practical physical watermarking approach to detect replay attacks in a CPS,” *Journal of Process Control*, vol. 116, pp. 136-146, 2022, doi: 10.1016/j.jprocont.2022.06.002.

[23] C. Trapiello and V. Puig, “Optimal Finite-time Watermark Signal Design got Replay Attack Detection using Zonotopes,” *IFAC-PapersOnLine*, vol. 55, no. 6, pp. 292-297, 2022, doi: 10.1016/j.ifacol.2022.07.144.

[24] C. Trapiello and V. Puig, “A Zonotopic-Based Watermarking Design to Detect Replay Attacks,” *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 11, pp. 1924-1938, 2022, doi: 10.1109/JAS.2022.105944.

[25] R. Goyal, C. Somarakis, E. Noorani and S. Rane, “Co-Design of Watermarking and Robust Control for Security in Cyber-Physical Systems,” *61st IEEE Conference on Decision and Control,* 2022, doi: 10.1109/CDC51059.2022.9992339.

[26] X. Q. Xie, H. H. Zhou and Y. H. Jin, “Strong tracking filter based adaptive generic model control,” *Journal of Process Control*, vol. 9, pp. 337-350, 1999, doi: 10.1016/S0959-1524(98)00052-3.

[27] R. Smith, “A decoupled feedback structure for covertly appropriating networked control systems,” *Proceedings of the 18th IFAC World Congress*, Milano, ITlau, pp. 90-95, 2011, doi: 10.3182/20110828-6-IT-1002.01721.

[28] A. Teixeira, D. Perez, H. Sandberg and K. H. Johansson, “Attack models and scenarios for networked control systems,” *Proceedings of the 1st International Conference on High Confidence Networked Systems*, Beijing, China, pp. 55-64, 2012, doi: 10.1145/2185505.2185515.

[29] F. Miao, M. Pajic and G. J. Pappas, “Stochastic game approach for replay attack detection,” 52nd IEEE Conference on Decision and Control, Florence, Italy, pp. 1854-1859, 2013, doi: 10.1109/CDC.2013.6760152.

[30] Y. Mo and B. Sinopoli, "Secure control against replay attacks," *Proceedings of the 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 911-918, 2009, doi: 10.1109/ALLERTON.2009.5394956.

[31] A. Khazraei, H. Kebriaei and F. R. Salmasi, "Replay attack detection in a multi agent system using stability analysis and loss effective watermarking," *American Control Conference,* 2017, doi: 10.23919/ACC.2017.7963694.

[32] R. M. G. Ferrari and A. M. H. Teixeira, "Detection and isolation of replay attacks through sensor watermarking," *Proceedings of the IFAC World Congress*, Toulouse, France, pp. 7363-7368, 2017, doi: 10.1016/j.ifacol.2017.08.1502.

[33] C. Bhowmick and S. Jagannathan, "Detection of Sensor Attacks in Uncertain Stochastic Linear Systems," *IEEE Conference on Control Technology and Applications*, pp. 706-711, 2019 doi: 10.1109/CCTA.2019.8920410.

[34] C. Bhowmick and S. Jagannathan, "Detection and Mitigation of Attacks in Nonlinear Stochastic System Using Modified $\chi^2$ Detector," *58th IEEE Conference on Decision and Control*, pp. 139-144, 2019, doi: 10.1109/CDC40024.2019.9029553.

[35] H. S. Sanchez, D. Rotondo, T. Escobet, V. Puig, J. Saludes and J. Quevedo, "Detection of replay attacks in cyber-physical systems using a frequency-based signature," *Journal of the Franklin Institute*, vol. 356, no. 5, pp. 2798-2824, 2019, doi: 10.1016/j.jfranklin.2019.01.005.

[36] C. Trapiello and V. Puig, "Replay attack detection using a zonotopic KF and LQ approach," *IEEE International Conference on Systems, Man and Cybernetics*, 2020, doi: 10.1109/SMC42975.2020.9282865.

[37] M. Porter, S. Dey, A. Joshi, P. Hespanhol, A. Aswani, M. J. Roberson and R. Vasudevan, "Detecting Deception Attacks on Autonomous Vehicles via Linear Time-Varying Dynamic Watermarking," *IEEE Conference on Control Technology and Applications,* Montreal, Canada, 2020, doi: 10.1109/CCTA41146.2020.9206278.

[38] M. Porter, P. Hespanhol, A. Aswani, M. J. Roberson and R. Vasudevan, "Detecting Generalized Replay Attacks via Time-Varying Dynamic Watermarking," *IEEE Transactions on Automatic Control*, vol. 66, no. 8, pp. 3502-17, 2021, doi: 10.1109/TAC.2020.3022756.

[39] L. Wu, D. Du, C. Zhang, M. Fei and I. Popovic, “An Active Detection Method for Generalized Replay Attacks Using Multiplicative Watermarking” *41st Chinese Control Conference,* Heifei, China, 2022, doi: 10.23919/CCC55666.2022.9901662.

[40] W. Li, H. Qian, M. Zhang, S. Wang, F. Wang and X. Zhu, “Stealthy replay attack detection of 3-DOF helicopter benchmark system using dynamic watermarking approach,” *Transactions of the Institute of Measurement and Control - Advances in measurement and control for unmanned systems*, pp. 1-11, 2022, doi: 10.1177/01423312221134326.

[41] R. Zhang, “Watermarking-based Discrete LQG Systems for Detecting Replay Attacks,” *35th Chinese Control and Decision Conference*, 2023, doi: 10.1109/CCDC58219.2023.10326524.

[42] A. Naha, A. Teixeira, A. Ahlen and S. Dey, “Sequential Detection of Replay Attacks,” *IEEE Transactions on Automatic Control*, vol. 68, no. 3, pp. 1941-48, 2023, doi: 10.1109/TAC.2022.3174004.

[43] A. Naha, A. Teixeira, A. Ahlen and S. Dey, “Sequential Detection of Replay Attacks with a Parsimonious Watermarking Policy,” *American Control Conference,* 2022, doi: 10.23919/ACC53348.2022.9867703.

[44] A. Ghamarilangroudi, S. Hashtrudi Zad and Y. Zhang, "Replay attack detection using switching multi-sine watermarking," *Proceedings of the 33rd Mediterranean Conference on Control and Automation (MED' 2025)*, Tangier, Morocco, June 2025.

[45] S. M. Kay, *Modern Spectral Estimation*, Prentice Hall, Englewood Cliffs, NJ, 1988.

[46] G. M. Jenkins and D. G. Watts, *Spectral Analysis and its Applications*, Holden-Day, San Francisco, 1968.

[47] T. T. Tran, O. S. Shin and J. H. Lee, "Detection of replay attacks in smart grid systems," *International Conference on Computing, Management and Telecommunications*, Ho Chi Minh City, Vietnam, pp. 298-302, 2013, doi: 10.1109/ComManTel.2013.6482409.

[48] M. Ma, P. Zhou, D. Du, C. Peng, M. Fei and H. M. AlBuflasa, “Detecting Replay Attacks in Power Systems: A Data-Driven Approach,” *Advanced Computational Methods in Energy, Power, Electric Vehicles, and Their Integration*, vol 763, 2017, doi: 10.1007/978-981-10-6364-0_45.

[49] P. Ramanan, D. Li and N. Gebraeel, "Blockchain-Based Decentralized Replay Attack Detection for Large-Scale Power Systems," in *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 8, pp. 4727-4739, 2022, doi: 10.1109/TSMC.2021.3104087.

[50] M. Bouslimani, F. B. S. Tayeb, Y. Amirat and M. Benbouzid, "Replay Attacks on Smart Grids: A Comprehensive Review on Countermeasures," *IECON 2024 - 50th Annual Conference of the IEEE Industrial Electronics Society*, Chicago, USA, 2024, pp. 1-6, doi: 10.1109/IECON55916.2024.10905194.

[51] H. R. Patel, "Replay Attack Detection in Smart Grids using Switching Multi-Sine Watermarking," M.S. thesis, Dept. of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada, 2023, https://spectrum.library.concordia.ca/id/eprint/992969/1/Patel_MASc_F2023.pdf.