GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

CONTROL DE VCS MEDIANTE SEÑALES NO AUDIBLES

Autor: Javier Valero Martí

Director: Javier Matanza Domingo

Co-Director: Gregorio López López

Madrid

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

CONTROL DE VCS MEDIANTE SEÑALES NO AUDIBLES

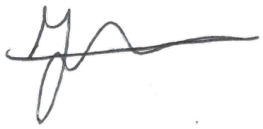
en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2021/2 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido

tomada de otros documentos está debidamente referenciada.



Fdo.: Javier Valero Martí Fecha: 7 / 7 / 2022

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Firmado por MATANZA
DOMINGO JAVIER - ***4397**
el día 07/07/2022 con un
certificado emitido por AC

Fdo.: Javier Matanza Domingo Fecha: 7 / 7 / 2022

LOPEZ LOPEZ
GREGORIO IGNACIO
- 75887558R

Digitally signed by LOPEZ
LOPEZ GREGORIO IGNACIO -
75887558R
Date: 2022.07.07 10:27:53
+02'00'

Fdo.: Gregorio López López Fecha: 7 / 7 / 2022

GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

CONTROL DE VCS MEDIANTE SEÑALES NO AUDIBLES

Autor: Javier Valero Martí

Director: Javier Matanza Domingo

Co-Director: Gregorio López López

Madrid

Agradecimientos

A mis profesores, por darme las herramientas para poder hacer este trabajo

A mis padres, por siempre apoyarme cuando más lo he necesitado

CONTROL DE VCS MEDIANTE SEÑALES NO AUDIBLES

Autor: Valero Martí, Javier
Director: Matanza Domingo, Javier
Co-Director: López López, Gregorio
Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

El proyecto se centra en el desarrollo de un modelo de ataque a dispositivos asistentes de voz basado en el empleo de señales no audibles. Este modelo se sustenta en la hipótesis de que la respuesta en frecuencia no lineal de los componentes del dispositivo atacado es una vulnerabilidad aprovechable para inyectar comandos de voz a partir de señales imperceptibles por un ser humano. Junto con el desarrollo del modelo, se han planteado una serie de pruebas experimentales para evaluar la viabilidad y efectividad del ataque, además de obtener información sobre el funcionamiento y la estructura interna de un asistente de voz.

Palabras clave: Asistente, voz, ataque, dispositivo, seguridad

1. Introducción

La industria de los dispositivos asistentes de voz representa un mercado que actualmente se encuentra en auge, según sugieren estudios recientes. Cada vez son más los usuarios que optan por adquirir un dispositivo con alguna herramienta de reconocimiento de voz instalada.

No obstante, como cualquier otro producto, estos dispositivos no se encuentran exentos de vulnerabilidades. En el año 2017, se publicó un documento científico que exponía una de estas vulnerabilidades y proponía un modelo de ataque para explotarla: *DolphinAttacks: Inaudible Voice Commands*. Los investigadores encargados de redactar este estudio pusieron de manifiesto que la respuesta en frecuencia no lineal que presentaban los micrófonos de estos dispositivos cuando las señales superaban los 20 kHz de frecuencia podía generar réplicas de la señal original en banda base.

Los resultados de dicho estudio demostraron que, si el espectro en frecuencia de las señales generadas en banda base es lo suficientemente similar al de un comando de voz reconocido por el sistema, se puede llegar a instruir al dispositivo para ejecutar acciones sin alertar al usuario. De llegar a materializarse, este ataque puede tener consecuencias realmente negativas.

2. Definición del proyecto

Este proyecto busca la consecución de tres objetivos principales:

- En primer lugar, se propondrá un modelo de ataque basado en el modelo *DolphinAttack*, aunque realizando una serie de modificaciones que exploran posibles escenarios en los que el nivel de riesgo de la amenaza se incrementaría.
- En segundo lugar, se probará el modelo de ataque diseñado para probar su viabilidad y efectividad. La batería de pruebas diseñada para ello tratará de evaluar la influencia

de diferentes factores, como podría ser la distancia entre emisor y receptor, en la consecución del ataque.

- En tercer lugar, se aprovechará el resultado de las pruebas realizadas para recabar información sobre el funcionamiento y la estructura interna de los dispositivos asistentes de voz.

3. Descripción del modelo de ataque

El modelo de ataque propuesto consta de los siguientes módulos:

- **Generador de señal**: las señales de voz utilizadas serán captadas por un dispositivo grabador de audio y almacenadas en formato digital.
- **Herramienta software para efectuar la modulación**: empleando el *software* de Matlab, se ha elaborado una herramienta que permite modular una señal de voz en banda base empleando diferentes técnicas de modulación en amplitud, tanto de doble banda lateral como de banda lateral única.
- **Emisor de señal:** para emitir la señal modulada, se ha optado para emplear un par de altavoces de carácter comercial, los cuales aseguran una respuesta en frecuencia lineal hasta los 20 kHz.
- **Dispositivo objetivo**: como dispositivo objetivo del ataque se emplearán diferentes equipos con distintas herramientas de reconocimiento de voz instaladas, a fin de comparar los resultados obtenidos para cada uno de ellos.

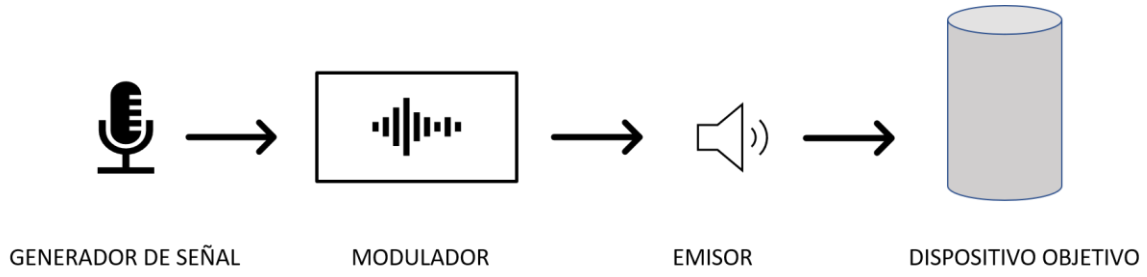


Ilustración 1. Esquema del modelo de ataque desarrollado

4. Resultados

Los resultados obtenidos tras la realización de las pruebas ponen de manifiesto las siguientes cuestiones:

- Los sistemas asistentes de voz cuentan con un umbral mínimo de potencia que debe superar toda señal de audio para ser interpretada por el *software* de reconocimiento de discurso interno al sistema. Este valor umbral es variable y depende del tipo de comando que el sistema espere recibir en cada momento.
- El uso de auriculares como entrada de audio del dispositivo reduce la máxima desviación en frecuencia de una señal en banda base tolerada por dicho dispositivo.
- Es inviable reproducir el modelo de ataque diseñado con sistemas emisores de audio que no aseguren una característica lineal de respuesta en frecuencia por encima de 20 kHz.

- Los altavoces de carácter comercial pueden llegar a reproducir señales audibles tomando como entrada señales no audibles. Estas señales reproducidas pueden ser entendidas por un asistente de voz.

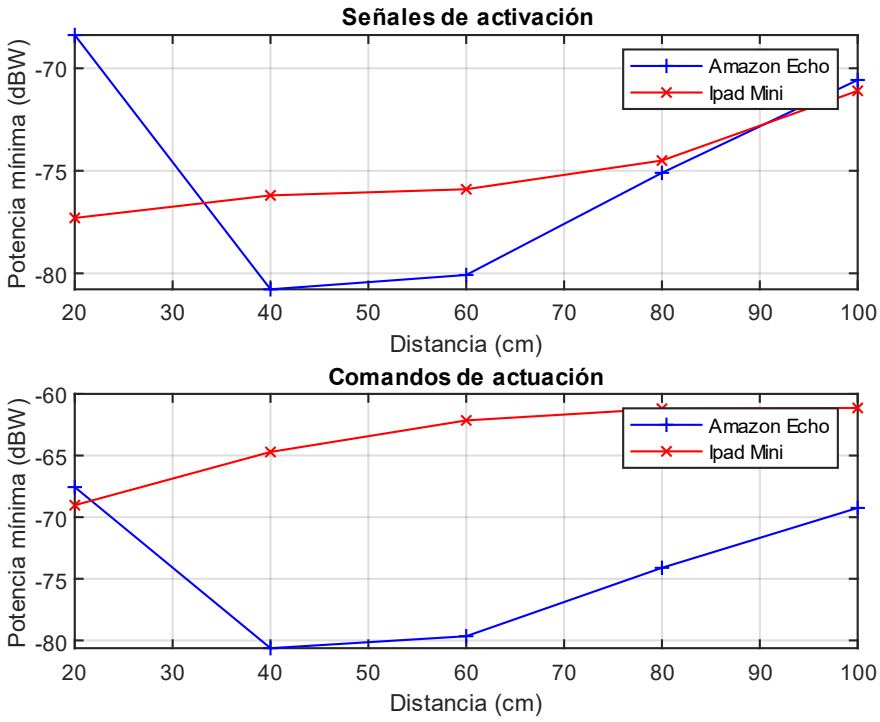


Ilustración 2. Representación gráfica de los resultados obtenidos en la prueba de umbral de potencia.

5. Conclusiones

A partir de los resultados expuestos, es posible elaborar las siguientes conclusiones:

- Los dispositivos que requieran de señal de activación son más seguros frente a modelos de ataque como el planteado.
- El uso de auriculares como entrada de audio puede servir como medida defensiva frente a modelos de ataque como el planteado.
- Los altavoces de carácter comercial presentan una característica no lineal que no permite su utilización en modelos de ataque como el planteado. No obstante, podrían ser empleados para modelar y estudiar la respuesta no lineal de los dispositivos asistentes de voz.

6. Referencias

Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., . . . Zhou, W. (2016).

Esteves, C. K. (2015). *IEMI Threats for Information Security: Remote Command Injection.* Paris.

Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., & Xu, W. (2017). *DolphinAttack: Inaudible Voice Commands.* Dallas, Tx.

VCS CONTROL USING NON-AUDIBLE SIGNALS

Author: Valero Martí, Javier
Supervisor: Matanza Domingo, Javier
Co-Supervisor: López López, Gregorio
Collaborating Entity: ICAI – Universidad Pontificia Comillas

ABSTRACT

The project focuses on the development of an attack model for voice assistant devices based on the use of non-audible signals. This model is based on the hypothesis that the non-linear frequency response of the components of the device under attack is a vulnerability that can be exploited to inject voice commands from signals that are imperceptible to a human being. Along with the development of the model, a series of experimental tests have been proposed to evaluate the feasibility and effectiveness of the attack, as well as to obtain information about the operation and internal structure of a voice assistant.

Keywords: Assistant, voice, attack, device, security

1. Introduction

The voice assistant device industry represents a booming market, according to recent studies. More and more users are opting to purchase a device with a voice recognition tool installed.

However, like any other product, these devices are not exempt from vulnerabilities. In 2017, a scientific paper was published that exposed one of these vulnerabilities and proposed an attack model to exploit it: DolphinAttacks: Inaudible Voice Commands. The researchers in charge of writing this study showed that the nonlinear frequency response presented by the microphones of these devices when signals exceeded 20 kHz frequency could generate replicas of the original baseband signal.

The results of this study showed that, if the frequency spectrum of the baseband signals generated is sufficiently similar to that of a voice command recognized by the system, the device can be instructed to perform actions without alerting the user. If it materializes, this attack can have truly negative consequences.

2. Definition of the project

This project seeks to achieve three main objectives:

- First, an attack model based on the DolphinAttack model will be proposed, although making a series of modifications that explore possible scenarios in which the threat risk level would be increased.
- Secondly, the designed attack model will be tested to prove its feasibility and effectiveness. The battery of tests designed for this purpose will try to evaluate the influence of different factors, such as the distance between sender and receiver, on the success of the attack.

- Thirdly, the results of the tests carried out will be used to gather information on the operation and internal structure of voice assistant devices, a field of study yet to be developed due to the opacity of the manufacturers. The results obtained can be used to prevent future vulnerabilities or propose defense measures for the developed attack.

3. Description of the attack model

The proposed attack model consists of the following modules:

- **Signal generator**: the voice signals used will be captured by an audio recording device and stored in digital format.
- **Software tool to perform the modulation**: using MATLAB software, a tool has been developed to modulate a baseband voice signal using different amplitude modulation techniques, both double sideband and single sideband.
- **Signal emitter**: to emit the modulated signal, we have chosen to use a pair of commercial loudspeakers, which ensure a linear frequency response up to 20 kHz.
- **Target device**: different equipment with different speech recognition tools installed will be used as the target device for the attack, in order to compare the results obtained for each of them.

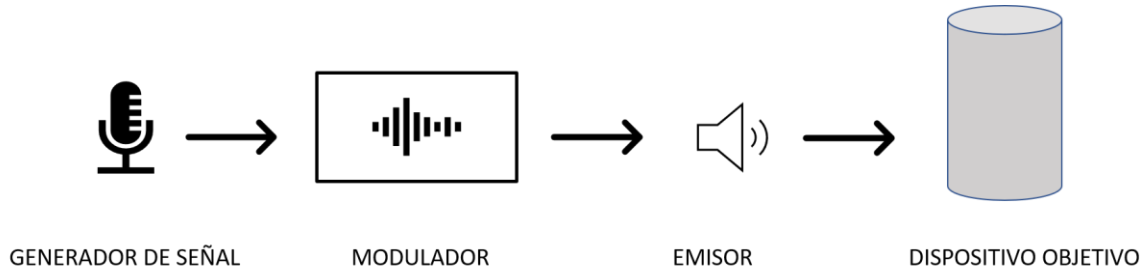


Ilustración 3. Schematic of the attack model developed

4. Results

The results obtained from the tests reveal the following issues:

- Voice assistant systems have a minimum power threshold that must be exceeded by any audio signal in order to be interpreted by the speech recognition software internal to the system. This threshold value is variable and depends on the type of command the system expects to receive at any given time.
- The use of headphones as the device's audio input reduces the maximum frequency deviation of a baseband signal tolerated by the device.
- It is unfeasible to reproduce the designed attack model with audio transmitter systems that do not ensure a linear frequency response characteristic above 20 kHz.

- Commercial loudspeakers can reproduce audible signals by taking non-audible signals as input. These reproduced signals can be understood by a voice assistant.

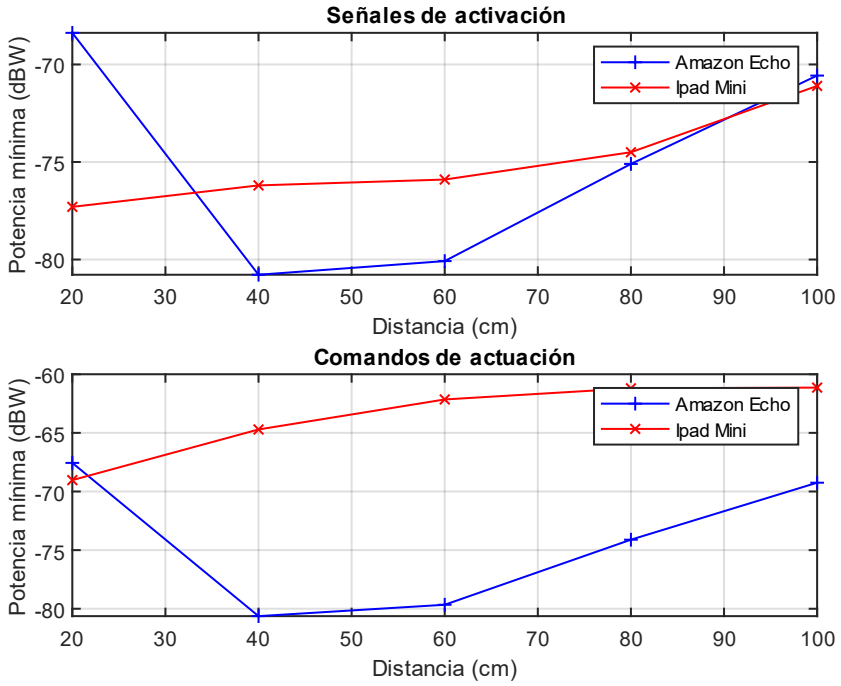


Ilustración 4. Graphical representation of the results obtained in the power threshold test.

5. Conclusions

From the above results, it is possible to draw the following conclusions:

- Devices that require activation signal are more secure against attack models such as the one posed.
- The use of headphones as audio input can serve as a defensive measure against attack models such as the one presented.
- Commercial loudspeakers present a non-linear characteristic that does not allow their use in attack models such as the one proposed. However, they could be used to model and study the nonlinear response of voice assistant devices.

6. References

Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., . . . Zhou, W. (2016).

Esteves, C. K. (2015). *IEMI Threats for Information Security: Remote Command Injection.* Paris.

Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., & Xu, W. (2017). *DolphinAttack: Inaudible Voice Commands.* Dallas, Tx.

Índice de la memoria

Capítulo 1. Introducción 7

1.1 Contexto del trabajo 7

1.2 Caracterización de la vulnerabilidad 8

1.3 Motivación 9

1.4 Descripción de los apartados del proyecto 10

Capítulo 2. Descripción de las Tecnologías 12

2.1 Funcionamiento de los sistemas asistentes de voz 12

2.1.1 Subsistema de captura de voz 13

2.1.2 Subsistema de reconocimiento de discurso 17

2.1.3 Subsistema de ejecución de comandos 20

2.2 Descripción de las técnicas de modulación empleadas 20

Capítulo 3. Estado de la Cuestión 26

3.1 Primeras vulnerabilidades encontradas 26

3.2 Ondas electromagnéticas como vector de ataque 27

3.3 Comandos de voz ocultos en señales audibles 29

3.4 Empleo de señales no audibles: backdoor attacks 31

3.5 Empleo de señales no audibles: dolphin attacks 33

3.6 Empleo de señales no audibles: surfing attacks 35

Capítulo 4. Definición del Trabajo 37

4.1 Justificación 37

4.1.1 Factores que incrementarían el riesgo de ataque 37

4.2 Objetivos 40

4.3 Metodología y planificación 41

Capítulo 5. Sistema Desarrollado 44

5.1 Hipótesis de trabajo 44

5.2 Diseño y desarrollo 47

5.2.1 Generador de señal 47

5.2.2 Modulador de señal 50

5.2.3 Método de filtrado 60

5.2.4 Sistema emisor de audio *63*
5.2.5 Dispositivo objetivo del ataque *63*

Capítulo 6. Análisis de Resultados ***65***
6.1 Diseño de pruebas 65
6.1.1 Prueba de umbral de potencia *65*
6.1.2 Prueba de desviación en frecuencia *67*
6.1.3 Prueba de efectividad del ataque *68*
6.2 Desarrollo de las pruebas y obtención de resultados 70
6.2.1 Prueba de umbral de potencia *71*
6.2.2 Prueba de desviación en frecuencia *78*
6.2.3 Prueba de reproducción de señales inaudibles *88*
6.2.4 Prueba de efectividad del ataque *90*
6.2.5 Pruebas adicionales *96*

Capítulo 7. Conclusiones y Trabajos Futuros ***97***
7.1 Resumen de las conclusiones obtenidas 97
7.1.1 Existencia de un valor umbral de potencia variable para la captación de señales por vcs *97*
7.1.2 Relación entre la arquitectura de un dispositivo asistente de voz y la potencia de las señales captadas *98*
7.1.3 Uso de auriculares como posible medida defensiva *99*
7.1.4 Imposibilidad de reproducir un ataque con señales no audibles haciendo uso de un altavoz de carácter comercial *100*
7.1.5 Posibilidad de utilizar altavoces para modelar la respuesta no lineal de un asistente de voz *101*
7.2 Trabajos futuros: Posibles medidas de defensa 102
7.2.1 Implementación de mecanismos de autenticación por voz *102*

Capítulo 8. Bibliografía ***104***

ANEXO I: ALINEACIÓN DEL PROYECTO CON LOS ODS ***107***

ANEXO II. HERRAMIENTA SOFTWARE PARA MODULACIÓN DE SEÑALES .. ***109***

Índice de figuras

Ilustración 1. Esquema del modelo de ataque desarrollado..... 9
Ilustración 2. Representación gráfica de los resultados obtenidos en la prueba de umbral de potencia..... 10
Ilustración 3. Schematic of the attack model developed 12
Ilustración 4. Graphical representation of the results obtained in the power threshold test.13
Ilustración 5. Subsistemas de un dispositivo asistente de voz..... 12
Ilustración 6. Estructura de micrófonos ECM y MEMS (Zhang, y otros, 2017). 14
Ilustración 7. Diagrama de bloques de un Conversor Analógico/Digital..... 15
Ilustración 8. Componentes del subsistema de captación de audio 16
Ilustración 9. Ejemplo de señal moduladora en el dominio del tiempo 21
Ilustración 10. Ejemplo de señal portadora en el dominio del tiempo 22
Ilustración 11. Ejemplo de señal modulada en amplitud en el dominio del tiempo..... 23
Ilustración 12. Ejemplo de señal modulada en amplitud con μ=0.5 24
Ilustración 13. Ejemplo de señal modulada en amplitud con μ=1.1 24
Ilustración 14. Planificación del proyecto2022 43
Ilustración 15. Efecto de una respuesta en frecuencia no lineal sobre un tono puro modulado en amplitud 46
Ilustración 16. Esquema del modelo de ataque desarrollado..... 47
Ilustración 17. Comparación entre una señal de voz natural y una artificial en el dominio temporal..... 48
Ilustración 18. Comparación entre una señal de voz natural y una artificial en el dominio de la frecuencia..... 49
Ilustración 19. Módulo del espectro en frecuencia de una señal de voz en banda base 51
Ilustración 20. Módulo del espectro en frecuencia de una señal de voz modulada en doble banda lateral con coseno..... 53
Ilustración 21. Módulo del espectro en frecuencia de una señal de voz modulada en doble banda lateral simple 55

Ilustración 22. Módulo del espectro en frecuencia de una señal de voz modulada en doble banda lateral simple 57
Ilustración 23. Respuesta en frecuencia de un filtro FIR de orden 14 (línea sólida) y un filtro IIR de orden 4 (línea punteada). Fuente: (Advsolned, 2020) 62
Ilustración 24. Medida de intensidad del ruido ambiente en el entorno de pruebas..... 70
Ilustración 25. Representación gráfica de los resultados obtenidos en la prueba de umbral de potencia. Relación entre potencia y distancia. 74
Ilustración 26. Influencia de la morfología del dispositivo en el resultado de la prueba de umbral de potencia..... 76
Ilustración 27. Representación gráfica de los resultados obtenidos en la prueba de umbral de potencia. Relación entre potencia y tipo de comando. 77
Ilustración 28. Representación gráfica de los resultados obtenidos en la prueba de desviación en frecuencia. Relación entre frecuencia y distancia. 84
Ilustración 29. Representación gráfica de los resultados obtenidos en la prueba de desviación en frecuencia según se usen o no auriculares 86
Ilustración 30. Modelo simplificado del aprovechamiento del efecto no linea 95
Ilustración 31. Vista general de la aplicación 109
Ilustración 32. Panel de selección de señal..... 110
Ilustración 33. Panel de selección de modulación 110
Ilustración 34. Panel de muestra, almacenaje y reproducción de señales 111

Índice de tablas

Tabla 1. Primeros 5 valores de los MFCC de una señal de voz natural y una artificial 50
Tabla 2. Dispositivos receptores empleados en la prueba del umbral de potencia 71
Tabla 3. Resultados de la prueba de umbral de potencia para el dispositivo Amazon Echo 72
Tabla 4. Resultados de la prueba de umbral de potencia para el dispositivo Ipad Mini 2 .. 73
Tabla 5. Dispositivos receptores empleados en la prueba de desviación de potencia 78
Tabla 6. Resultados de la prueba de desviación en frecuencia para el dispositivo Amazon Echo 80
Tabla 7. Resultados de la prueba de desviación en frecuencia para el dispositivo Ipad Mini 2 80
Tabla 8. Resultados de la prueba de desviación en frecuencia para el dispositivo Samsung Galaxy S21s / Bixby 81
Tabla 9. Resultados de la prueba de desviación en frecuencia para el dispositivo Samsung Galaxy S21s / Bixby usando auriculares 81
Tabla 10. Resultados de la prueba de desviación en frecuencia para el dispositivo Samsung Galaxy S21s / Google Assistant 82
Tabla 11. Resultados de la prueba de desviación en frecuencia para el dispositivo Samsung Galaxy S21s / Google Assistant usando auriculares 82
Tabla 12. Media aritmética y desviación típica de los resultados de la prueba de desviación en frecuencia para cada dispositivo asistente 83
Tabla 13. Valor medio de la desviación en frecuencia tolerada por cada dispositivo para cada una de las distancias probadas 85
Tabla 14. Comparación de los resultados obtenidos en la prueba de desviación en frecuencia según se usen o no auriculares 86
Tabla 15. Dispositivos empleados en la prueba de audibilidad 88
Tabla 16. Resultados obtenidos en la prueba de audibilidad para cada dispositivo emisor 89
Tabla 17 Resultados obtenidos en la prueba de inteligibilidad para cada dispositivo emisor 89

Tabla 18. Resultados obtenidos en la prueba de efectividad del ataque para el receptor Amazon Echo 91
Tabla 19. Resultados obtenidos en la prueba de efectividad del ataque para el receptor Ipad Mini 2 92
Tabla 20. Resultados obtenidos en la prueba de efectividad del ataque para el receptor Samsung Galaxy S21/ Bixby 92
Tabla 21. Resultados obtenidos en la prueba de efectividad del ataque para el receptor Samsung Galaxy S21/ Google Assistant 92
Tabla 22. Resultados obtenidos en la prueba de efectividad del ataque para el emisor SBS 260 93
Tabla 23. Resultados obtenidos en la prueba de efectividad del ataque para el emisor Z120 93
Tabla 24. Resultados de prueba del modelo DolphinAttack sobre un dispositivo Amazon Echo. Fuente: (Zhang, y otros, 2017) 94

Capítulo 1. INTRODUCCIÓN

1.1 CONTEXTO DEL TRABAJO

Los dispositivos inteligentes para uso en el hogar constituyen un mercado que ha experimentado un crecimiento significativo en los últimos años, motivado fundamentalmente por el desarrollo de nuevas tecnologías de red, como 5G o WiFi-6, y el estudio de posibles implementaciones de un sistema tan ambicioso como el que plantea el paradigma de *Internet of Things (IoT).* Según estudios recientes publicados por la compañía *GlobalData*, para el año 2025 estos dispositivos habrán constituido un mercado que moverá 75.000 millones de dólares (Computerworld, 2019).

Dentro de este mercado, uno de los pilares fundamentales son los asistentes personales controlados por voz o VCSs[1], los cuales se definen como una serie de dispositivos capaces de reconocer e interpretar instrucciones recitadas de forma oral por un usuario para ejecutar determinadas acciones en respuesta. El crecimiento en popularidad de estos sistemas entre los usuarios va inevitablemente ligado al desarrollo del mercado de sistemas inteligentes para el hogar y no parece que esta tendencia ascendente vaya a experimentar cambio alguno en un futuro cercano. Así lo indica el informe *"Global Speech and Voice Recognition Market-Industry Trends and Forecast 2025*" (GlobalInfoResearch, 2020), el cual prevé un crecimiento del 25.7% por año, entre 2018 y 2025, para los sistemas que aplican tecnología de reconocimiento de voz.

Sin embargo, estos sistemas no están exentos de vulnerabilidades que, de llegar a ser explotadas, podrían traducirse en un impacto sustancialmente negativo para los usuarios. Este hecho se puso por primera vez de manifiesto en un artículo de investigación titulado *DolphinAttacks: Inaudible Voice Commands,* en el que se estudia la viabilidad de introducir

[1] Acrónimo procedente del término en inglés: "*Voice Controlled Systems*"

señales de voz, por medio de un proceso de modulación, en señales no audibles que permitan controlar un asistente de voz sin que las instrucciones solicitadas sean percibidas por los usuarios. Los resultados del estudio ponen de manifiesto la viabilidad del ataque sobre un amplio grupo de dispositivos de control por voz, entre ellos algunos de los líderes del mercado, como el asistente Siri, instalado en dispositivos móviles y tabletas de Apple, o el asistente Alexa, propio de dispositivos comercializados por Amazon.

Este es el contexto en el que se enmarca el trabajo a desarrollar: la existencia de una vulnerabilidad potencialmente aprovechable en un mercado en auge, el cual tiene previsto englobar cada vez a un número más alto de usuarios.

1.2 CARACTERIZACIÓN DE LA VULNERABILIDAD

Toda vulnerabilidad presente en un sistema o dispositivo electrónico tiene asociado un nivel de riesgo, el cual puede emplearse para representar el grado de urgencia con el que dicha vulnerabilidad debería ser subsanada. Para caracterizar un nivel de riesgo necesitamos tres factores: la existencia de una vulnerabilidad, la probabilidad de que dicha vulnerabilidad llegue a ser explotada y el impacto que esto supondría en aquello que queremos proteger, en este caso, la privacidad de los usuarios.

La existencia de una vulnerabilidad queda sobradamente probada a partir de los resultados aportados por el estudio referenciado en el apartado anterior. En el caso de aquellos ataques que emplean ondas de frecuencias no audibles, esta vulnerabilidad está asociada a la característica no lineal presentada por los micrófonos que, al menos hasta el momento del ataque, los principales fabricantes de dispositivos asistentes de voz utilizaban para manufacturar sus productos.

La posibilidad de explotar esta vulnerabilidad también se manifiesta en el Estado del Arte, pues tal como se expone en los puntos 3.5 y 3.6, los modelos de ataque *DolphinAttack* y *SurfingAttack* consiguen aprovechar dicha vulnerabilidad para inducir la ejecución de

comandos de voz en un dispositivo. Por tanto, cualquier potencial atacante en posesión del material adecuado estaría en disposición de replicar el procedimiento.

El último factor clave para caracterizar el nivel de riesgo es, por tanto, el impacto. Este factor trata de ser una forma de cuantificar, o al menos de describir, las consecuencias a las que debería hacer frente un usuario el cual se viese afectado por un modelo de ataque como los anteriormente referenciados. Alguno de los posibles efectos negativos que podrían afectar al usuario del dispositivo atacado serían:

- **Espionaje**: Con los métodos descritos, se podría instruir al dispositivo atacado para iniciar una llamada de audio o vídeo con un dispositivo manejado por el atacante. De esta manera, se podría tener acceso a la imagen o sonido del espacio colindante al dispositivo, un espacio generalmente privado, sin alertar al usuario.
- **Denegación de servicio**: Sería posible, por medio de comandos, conseguir que un dispositivo desactive los servicios de conexión a la red o recepción de llamadas (por ejemplo, mediante la activación del "modo avión"), sin que el usuario sea consciente de ello.
- **Infección por *malware***: Otra posibilidad sería inducir al dispositivo atacado a realizar una búsqueda por Internet y visitar un sitio web malicioso, con algún tipo de *malware* desarrollado por el atacante, para intentar forzar la infección del dispositivo.
- **Publicación de información falsa / Difamación**: En algunos dispositivos, es posible controlar, por medio de comandos de voz, el envío de mensajes de texto o publicación de contenido en redes sociales. Por tanto, sería viable que un atacante utilizase los métodos anteriormente descritos para forzar a un dispositivo a enviar algún tipo de mensaje, en nombre del usuario, sin la autorización de este.

1.3 MOTIVACIÓN

La motivación para realizar este proyecto surge de la necesidad de obtener más información acerca del funcionamiento de los sistemas asistentes de voz, a fin de caracterizar de manera más precisa el nivel de riesgo al que están sometidos actualmente.

Es cierto que la opacidad de las empresas fabricantes de estos productos, en lo relativo a la arquitectura subyacente a ellos y los componentes empleados en el proceso de fabricación, limita la información con la que cuenta un potencial atacante a la hora de buscar vulnerabilidades. No obstante, cuando una vulnerabilidad que afecta a alguno de estos dispositivos se da a conocer, resulta complicado que investigadores externos a la industria puedan ofrecer algún tipo de respuesta.

Por ello, el desarrollo de este trabajo trata, por medio de pruebas experimentales, de obtener resultados que permitan entender mejor los fundamentos del funcionamiento de estos dispositivos. De esta manera, podría ser más sencillo predecir futuras vulnerabilidades y ofrecer soluciones a las ya conocidas.

Asimismo, también existe interés por estudiar la evolución de la industria de fabricación de VCSs, a fin de determinar si las nuevas generaciones de dispositivos son más robustas que las anteriores frente a ataques ya conocidos. Para ello, durante el desarrollo del trabajo se tratará de reproducir modelos de ataque anteriormente probados contra dispositivos que se encuentran actualmente en el mercado.

1.4 DESCRIPCIÓN DE LOS APARTADOS DEL PROYECTO

El presente documento consta de los siguientes apartados, además del actual:

- En el **Capítulo 2. Descripción de las Tecnologías** se realiza una exposición de los fundamentos de dos conceptos clave para entender el contenido del trabajo: sistemas asistentes de voz y modulación de señales en amplitud.
- El **Capítulo 3. Estado de la Cuestión** está dedicado a poner en contexto el trabajo realizado, explicando el alcance de estudios anteriores relacionados con la temática central de este proyecto.
- En el **Capítulo 4. Definición del Trabajo** se realiza una descripción formal del trabajo a realizar, ofreciendo una justificación para el mismo y estableciendo objetivos a cumplir y una metodología a seguir.

- El **Capítulo 5. Sistema Desarrollado** está dedicado a exponer el modelo de ataque desarrollado. Se plantea la hipótesis central sobre la que se sustenta el proyecto y se explica en detalle cada una de las diferentes partes de este modelo.
- En el **Capítulo 6. Análisis de Resultados**, se proponen una serie de pruebas para evaluar la viabilidad del modelo propuesto. Tras la realización de las mismas, se describen y comentan los resultados.
- El **Capítulo 7. Conclusiones y Trabajos Futuros** sirve como conclusión del trabajo realizado, poniendo de manifiesto las principales implicaciones de los resultados obtenidos en el apartado anterior y proponiendo líneas de trabajo para futuros proyectos.
- Finalmente, en el **Capítulo 8. Bibliografía** se recogen las referencias bibliográficas que sirven como fuente de información a este trabajo.

El documento principal cuenta con dos documentos anexos:

- El **Anexo I** describe la alineación del proyecto con los Objetivos de Desarrollo Sostenible.
- El **Anexo II** sirve a modo de documentación de la herramienta *software* empleada para llevar a cabo los procesos de modulación de señal.

Capítulo 2. DESCRIPCIÓN DE LAS TECNOLOGÍAS

2.1 FUNCIONAMIENTO DE LOS SISTEMAS ASISTENTES DE VOZ

La arquitectura subyacente en los sistemas controlables por voz sigue siendo un campo con más espacio ocupado por teorías y propuestas que por información contrastada, debido a la opacidad que muestran los diferentes fabricantes a la hora de desvelar los entresijos técnicos de estas aplicaciones. La reticencia de las empresas a publicar información sobre el funcionamiento de los dispositivos manufacturados tiene su origen en una cuestión puramente económica: estas empresas pelean dentro de un mercado en régimen de competencia en el que cualquier ventaja técnica de la que disponga con respecto a los competidores puede marcar la diferencia. Además, esta posición se ha visto reforzada a la luz de las últimas vulnerabilidades descubiertas en estos sistemas, pues a mayor cantidad información disponible para un posible atacante, mayor posibilidad de que nuevas vulnerabilidades sean descubiertas.

No obstante, a partir de su funcionamiento y teniendo en cuenta la estructura de aplicaciones con capacidades similares, es posible inferir que todos los sistemas asistentes de voz están compuestos, como mínimo, por los subsistemas indicados en la Ilustración 5: subsistema de captura de voz, subsistema de reconocimiento de discurso y subsistema de ejecución de comandos.

Ilustración 5. Subsistemas de un dispositivo asistente de voz

2.1.1 SUBSISTEMA DE CAPTURA DE VOZ

La tarea general de este subsistema es la de captar el sonido ambiente en forma de ondas acústicas y transformarlo en señales eléctricas que serán procesadas en la fase de reconocimiento de discurso. Para ello, son esenciales dos elementos:

2.1.1.1 Transductor

Es un dispositivo capaz de convertir una forma de energía en otra. Si transforma la presión generada por las ondas acústicas sobre el aire en energía eléctrica se denomina transductor electroacústico o, comúnmente, micrófono (Brüel & Kjær, s.f.). Una de las maneras más comunes de lograr esta conversión es haciendo uso de condensadores: la presión del aire ocasiona deformaciones físicas en el condensador que a su vez se traducen en un cambio de capacidad y, por tanto, de voltaje a través del mismo.

La operación de conversión generalmente presenta una expresión lineal del tipo $y = \alpha x$, donde α representa la sensibilidad del transductor. No obstante, la realidad es que dicha característica lineal en los micrófonos únicamente se manifiesta para un determinado rango de frecuencias. A partir de este rango, las señales experimentan transformaciones no lineales que ya no pueden ser modeladas de manera sencilla como en el caso anterior.

Dentro de los micrófonos que hacen uso de condensadores encontramos dos tipos principales: *ECM*[2] y *MEMS*[3]. En la Ilustración 6 se puede observar cómo, en ambos diseños, se aprovecha el cambio de capacidad en uno o varios condensadores para llevar a cabo la transformación energética que genera una señal eléctrica de valor continuo. No obstante, las diferencias presentes en los componentes y estructura interna de estos diseños ocasionan distinciones que los fabricantes deben tener en cuenta a la hora de elegir un tipo u otro de transductor para emplear en sus dispositivos.

[2] Acrónimo de *"Electret Condenser Microphone". En castellano, "Micrófono de condensador electrolítico".*
[3] Acrónimo de *"Micro-Electro-Mechanical System". En castellano, "Sistema micro-electromecánico"*

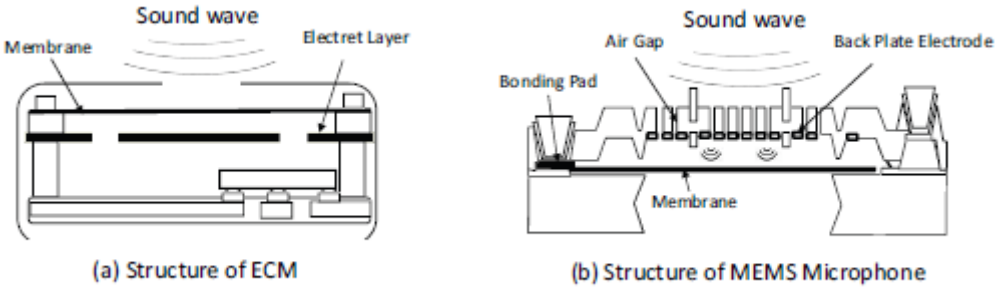


Ilustración 6. Estructura de micrófonos ECM y MEMS (Zhang, y otros, 2017).

Por un lado, los micrófonos de tipo *MEMS* presentan un tamaño más reducido que los de tipo *ECM*, lo cual supone una ventaja aprovechable en la construcción de dispositivos sujetos a una restricción o limitación de espacio. Además, resultan especialmente útiles para sistemas destinados a trabajar en ambientes con elevado ruido electromagnético debido a su baja impedancia y su coste es generalmente menor que el de los diseños de tipo *ECM* (Rose, 2019).

Por otro lado, los *ECM* representan un diseño con más tiempo en el mercado y que, por lo tanto, ha sido empleado en una enorme cantidad de sistemas, por lo que mantener esta tecnología en actualizaciones poco complejas de los mismos podría ser una solución interesante para los fabricantes. Además, el hecho de presentar un volumen superior a los *MEMS* hace posible desarrollar en estos diseños soluciones que permiten alcanzar características como la impermeabilidad, las cuales pueden resultar esenciales en entornos de trabajo concretos.

No obstante, las ventajas en cuanto a tamaño y coste de los micrófonos *MEMS* les sitúan como la solución mayormente adoptada por la industria tecnológica a la hora de desarrollar aplicaciones que requieran de un subsistema de captura de audio (Zhang, y otros, 2017).

2.1.1.2 Conversor Analógico/Digital

Es la unidad encargada de convertir las señales eléctricas generadas por el transductor del dominio analógico al dominio digital. La digitalización de las señales de audio es necesaria para facilitar su transmisión y procesamiento a través de las unidades que componen el subsistema de reconocimiento de discurso. Este proceso requiere de diferentes fases de tratamiento de la señal por la cual esta será discretizada en los ejes de amplitud y tiempo, tal como se muestra en la Ilustración 7.

- La discretización en el eje temporal se consigue tomando muestras del valor de amplitud de la señal en intervalos equiespaciados de tiempo, definidos por la frecuencia de muestreo del sistema.
- La discretización en el eje de amplitud se consigue comparando los valores de las muestras tomadas previamente con una escala predeterminada de valores equiespaciados y separados por el denominado "escalón de cuantización". Cada valor original será sustituido por el valor más próximo dentro de esta escala. Finalmente, este valor será representado en términos digitales por el número de bits que caractericen la salida del sistema.

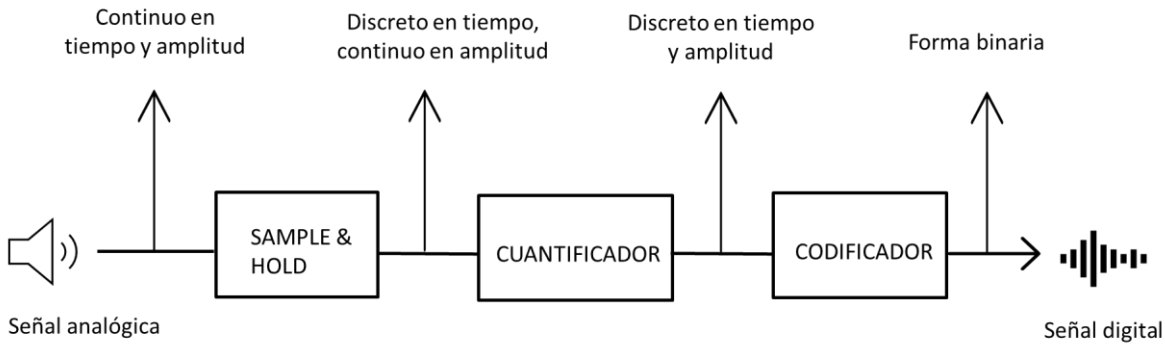


Ilustración 7. Diagrama de bloques de un Conversor Analógico/Digital

La representación digital de la señal no mantiene una equivalencia exacta con la forma de señal en el domino analógico, de manera que se debe llegar a un compromiso entre la precisión ofrecida por el conversor y la requerida por el resto de unidades del sistema para desempeñar un correcto funcionamiento. En términos del conversor, la precisión del mismo puede verse incrementada aumentando la frecuencia de muestreo o el número de niveles de cuantización empleados (por tanto, el número de bits necesarios para representar la salida), aunque esta mejora de prestaciones conllevaría a su vez un aumento en la complejidad del *hardware* requerido y, por tanto, del coste de la unidad.

La frecuencia de muestreo utilizada por los conversores empleados en la industria oscila entre 8 y 768 kHz, aunque los valores más empleados de facto son 16 kHz y 48 kHz (Texas Instruments, 2019). En cuanto al número de bits empleados para representar la salida, las opciones más empleadas son 16, 20, 24 o 32.

2.1.1.3 Componentes adicionales

La funcionalidad de las dos unidades principales del subsistema, transductor y conversor, se ve generalmente complementada con otros dos elementos situados entre ambos: amplificador y filtro. La Ilustración 8 muestra todos los componentes del subsistema:

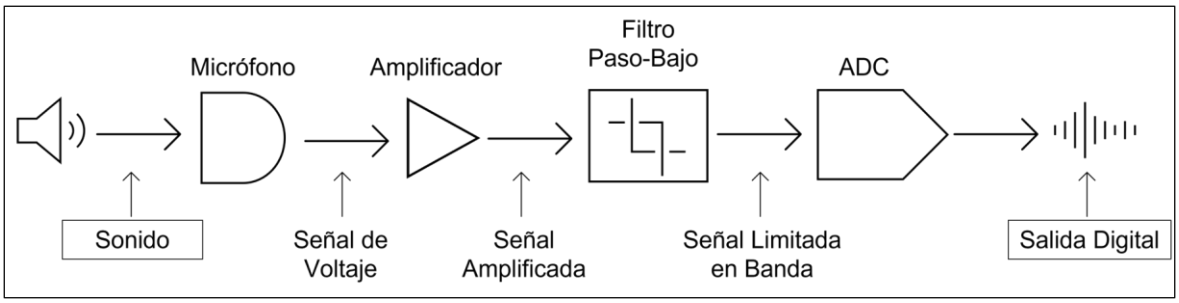


Ilustración 8. Componentes del subsistema de captación de audio

Amplificador de potencia: empleado para aumentar el nivel de potencia de la señal de manera que los valores de amplitud de la misma se ajusten a la escala determinada por los escalones de cuantización del conversor analógico/digital.

Filtro paso bajo: empleado para evitar el solapamiento espectral de la señal tras su paso por el conversor. Según el criterio de Nyquist, el ancho de banda máximo que puede presentar la señal a la entrada del conversor para asegurar que no se produce *aliasing*[4] es igual a la mitad del valor de frecuencia de muestreo empleado. Por ejemplo, para un valor de $f = 16\ kHz$, la señal de entrada debería estar limitada en frecuencia a 8 kHz. Pasa asegurar que se cumple esta condición, se emplea un filtro paso bajo con frecuencia de corte igual al valor de frecuencia al que se desea limitar la señal de entrada al conversor.

Evitar el solapamiento espectral de la señal procesada por el subsistema de captación de audio es un requisito indispensable para los sistemas asistentes de voz, pues de lo contrario se produciría una pérdida de información de la señal que podría afectar al desempeño del sistema de reconocimiento de discurso, lo cual repercutiría negativamente en la experiencia de los usuarios del sistema.

2.1.2 SUBSISTEMA DE RECONOCIMIENTO DE DISCURSO

La función global del conjunto de procesos que constituyen este subsistema es la de procesar la señal recibida desde el sistema de captación de audio (en formato digital) y determinar si dicho conjunto de muestras equivale a una señal de voz que a su vez represente un comando válido dentro del alcance del sistema. Este proceso general puede a su vez dividirse en una serie de fases en las que la señal de entrada es sometida a diferentes tratamientos de procesamiento y análisis:

Fase previa al procesado: en esta fase, se aplica un filtrado rudimentario a las señales recibidas tratando de separar las muestras de voz del resto de señales de audio que hayan

[4] Término anglosajón referido al solapamiento en frecuencia entre las réplicas del espectro de una señal, generadas tras someter a la misma a un proceso de muestreo. Aparece debido al uso de un valor de frecuencia de muestreo demasiado bajo (Wilczek, 2019).

sido captadas por el sistema. Para ello, las señales son filtradas de nuevo en frecuencia, mediante una serie de filtros paso bajo con un valor de frecuencia de corte cercano a las máximas frecuencias de voz emisibles (entre 6 y 8 kHz). Las señales también son filtradas en función de su nivel de potencia, siendo descartadas todas aquellas muestras correspondientes a intervalos de tiempo donde no se supere el umbral mínimo de potencia fijado por los diseñadores del sistema.

Fase de obtención de características: En la siguiente fase del proceso, la información de señal se divide en tramos para facilitar su análisis en tiempo real. A cada tramo de muestras se le aplica un proceso de transformación por el cual se trata de extraer una serie de rasgos característicos de dicho fragmento de señal que puedan ser utilizados para modelar la información de audio que este contiene.

El proceso más usado en esta fase por los sistemas asistentes de voz es el cálculo de los MFCC[5]. Esta práctica permite obtener los coeficientes de una transformación que busca emular la forma que tiene el oído humano de percibir los sonidos, realizando un análisis logarítmico a frecuencias altas y lineal a frecuencias bajas (Kaberpanthi & Datar, 2014). A nivel general, se trata de una transformación que trata de extraer de la señal las frecuencias que el oído humano percibiría con mayor nivel de potencia (frecuencias dominantes) como una forma de caracterizar el fragmento de audio que está siendo analizado.

Fase de predicción basada en un modelo: En esta etapa del proceso se comparan los valores extraídos de la señal en la fase anterior con un modelo definido previamente para tratar de predecir el fragmento de texto que pudiera estar contenido en el tramo de señal que está siendo analizado.

Dentro de los modelos predictivos empleados por las diferentes aplicaciones de reconocimiento de voz, se considera que uno de los más comunes es el uso de HMM[6], aunque

[5] Acrónimo de *Mel-frequency Cepstral Coefficients*
[6] Acrónimo de *Hidden Markov Models*

no se descarta el empleo de algoritmos de predicción basados en redes neuronales recurrentes (Zhang, y otros, 2017).

La aplicación de estos algoritmos de predicción permite generar un vector con diferentes opciones para representar el texto enunciado por la voz contenida en la parte de audio que está siendo objeto de análisis.

Finalmente, en la **fase de post-procesado**, se ordenan las diferentes opciones de texto ofrecidas por el modelo de predicción de menor a mayor probabilidad aplicando criterios extra que pueden influir en el contenido del discurso, como reglas gramaticales o el uso de expresiones propias de una determinada región.

Todos los procesos anteriores se llevan a cabo en tiempo real y de manera ininterrumpida mientras que el dispositivo en cuestión se encuentre en un estado de escucha activa. El estado del sistema vendrá dado por los requisitos de diseño del mismo, pudiendo diferenciar entre dos tipos de sistemas en función de si hacen uso de algún tipo de mecanismo de activación o no.

Los sistemas que no requieren ningún tipo de mecanismo de activación se encuentran siempre en fase de escucha activa, lo que implica que sus micrófonos están continuamente captando el sonido ambiente y las señales captadas están siendo procesadas por el sistema de reconocimiento de voz.

Por otra parte, si el sistema requiere de activación previa, se debe tener en cuenta si la activación se realiza por medio de comando de voz o de acción del usuario. En el caso de activación por medio de un comando de voz, el sistema se encontrará en un estado de escucha semiactiva: todas las señales captadas están siendo procesadas, pero los algoritmos predictivos únicamente comparan el texto recibido con el comando esperado. El resto de comandos se encuentran bloqueados hasta que el sistema es activado. Por otro lado, si el proceso de activación requiere de una acción ejecutada por parte del usuario (pulsar un botón o abrir una determinada aplicación), el sistema no procesará ninguna de las señales captadas por los micrófonos hasta que haya sido activado.

2.1.3 SUBSISTEMA DE EJECUCIÓN DE COMANDOS

Este subsistema entra en juego cuando los algoritmos de predicción de texto empleados en la fase de reconocimiento de discurso determinan que la señal de audio captada por los micrófonos del dispositivo contiene algún comando de voz que el sistema reconoce como válido. Una vez se da esta situación, el sistema se encarga de ejecutar la acción vinculada al comando que ha sido reconocido. Esta acción puede abarcar desde la modificación de parámetros de uso del dispositivo hasta la invocación de aplicaciones desarrolladas por terceros si el sistema cuenta con autorización para ello.

2.2 DESCRIPCIÓN DE LAS TÉCNICAS DE MODULACIÓN EMPLEADAS

En el desarrollo de este proyecto, es necesario trabajar con señales de voz situadas en una banda de frecuencias superior a la banda estándar de las señales de voz humana. Para poder desplazar en frecuencia estas señales sin perder información de las mismas, es necesario aplicar técnicas de modulación.

En el proceso de modulación intervienen las siguientes señales:

- **Señal moduladora:** es la señal que se desea desplazar en frecuencia para facilitar su transmisión. Contiene la información de señal original.
- **Señal portadora:** permite el transporte de la información contenida en la señal moduladora. Esta información será almacenada en esta señal a través de cambios en alguno de sus parámetros (amplitud, frecuencia o fase). La frecuencia base de la modulación viene dada por la frecuencia fundamental de esta señal.
- **Señal modulada:** Es el resultado de la modificación de la señal portadora en base a la moduladora. Es la señal que será transmitida por el canal de comunicación.

Dentro de las diferentes técnicas de modulación conocidas, en este proyecto se ha optado por el uso de métodos de modulación en amplitud.

Como su propio nombre indica, la modulación en amplitud o AM utiliza la amplitud de la señal portadora como parámetro variable para almacenar la información de la señal moduladora. El efecto de la modulación sobre la forma de onda de las señales se puede apreciar visualmente mediante una representación de la amplitud de las señales implicadas con respecto del tiempo.

Se toma como ejemplo la señal definida por la Ecuación 1 y representada en el dominio temporal en la Ilustración 9 :

Ecuación 1 $$m(t) = Am\,[cos(2\pi f_1 t) + cos(2\pi f_2 t)]\,, \qquad Am, f_1, f_2 \in \mathbb{N}$$

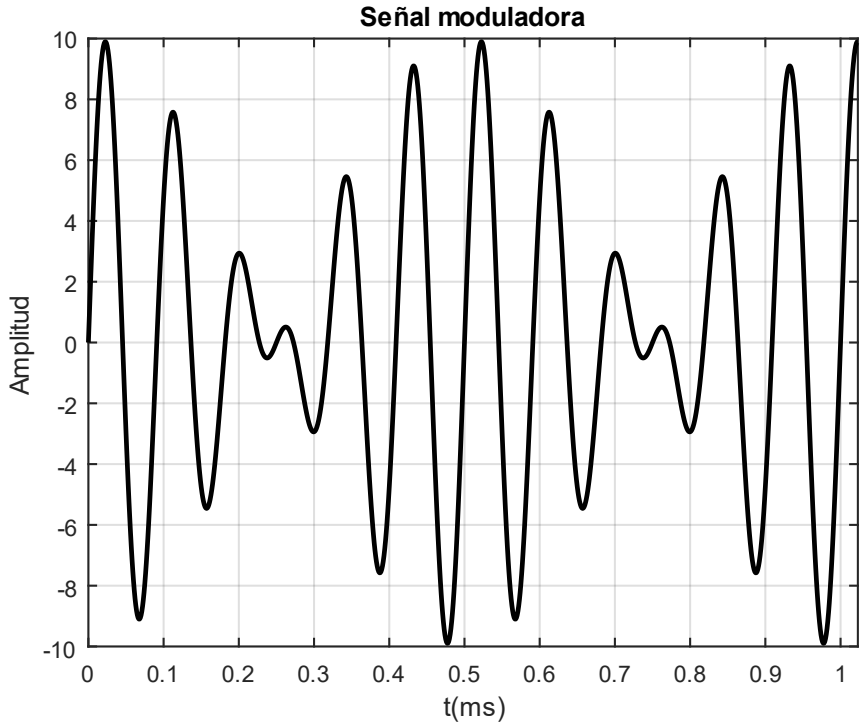


Ilustración 9. Ejemplo de señal moduladora en el dominio del tiempo

Como señal portadora, se toma un tono simple de frecuencia mayor que la frecuencia de la señal moduladora, definido por la Ecuación 2. Su representación en el dominio del tiempo se encuentra en la Ilustración 10.

Ecuación 2 $$p(t) = Ac\ cos(2\pi f_w t)\,, \quad Ac, f_w \in \mathbb{N}, \quad f_w \gg f_1, f_2$$

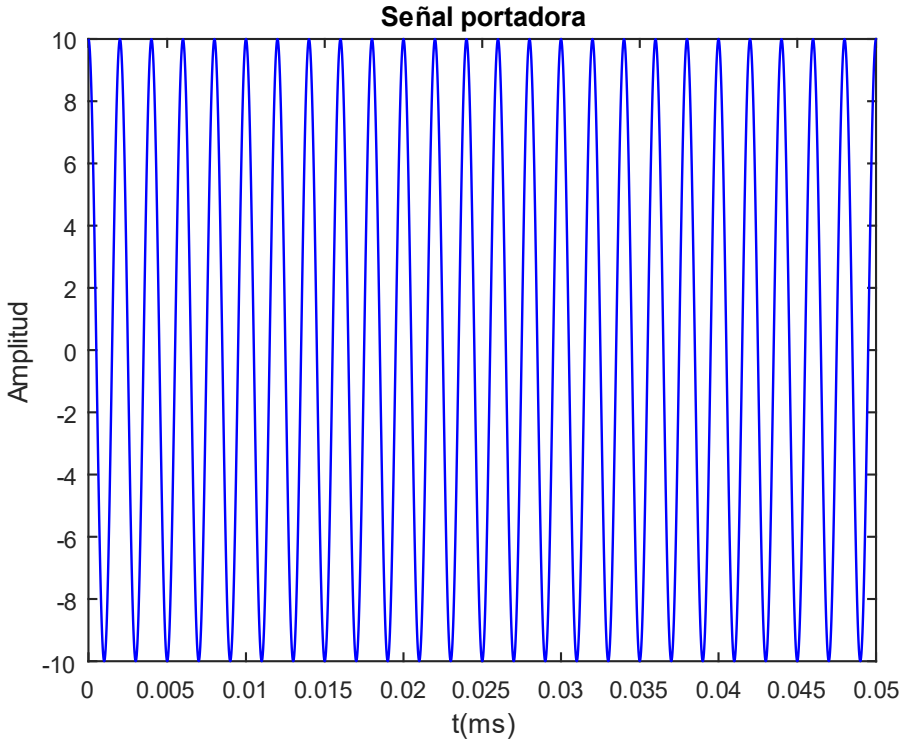


Ilustración 10. Ejemplo de señal portadora en el dominio del tiempo

La señal modulada será el resultado de adaptar la amplitud de la señal portadora a la forma de onda de la moduladora. Se puede lograr mediante la siguiente expresión:

Ecuación 3 $$y(t) = [1 + \mu\, m(t)]\, p(t)\,, \quad \mu \in \mathbb{N}$$

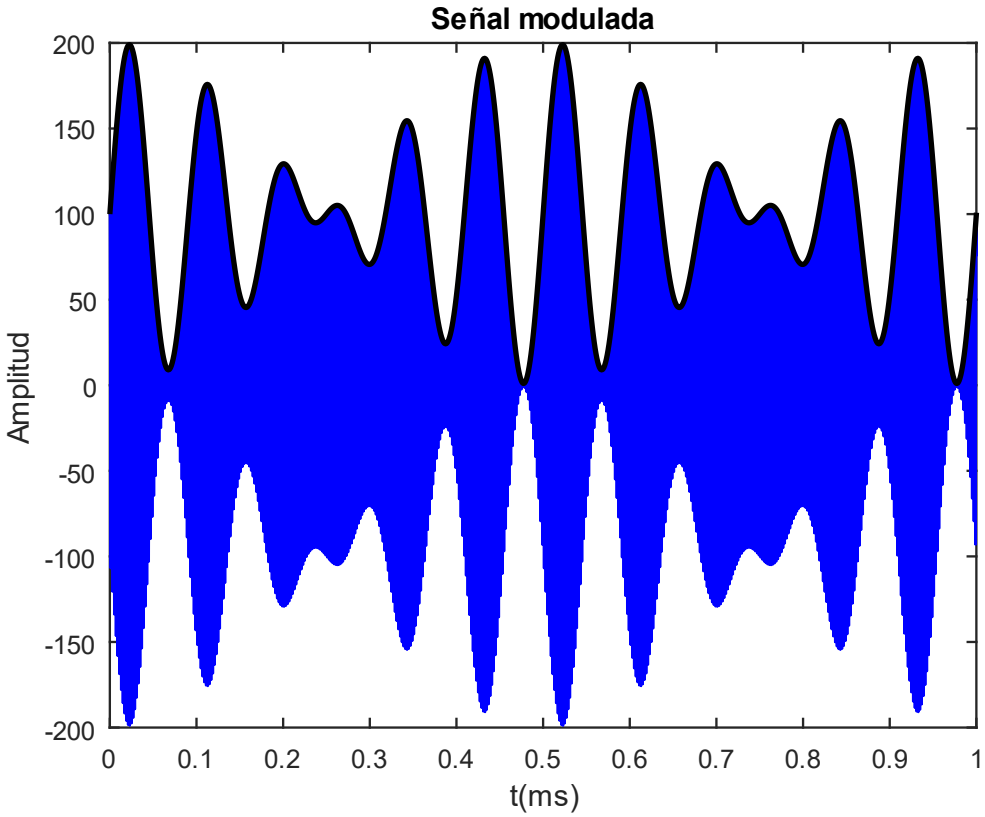


Ilustración 11. Ejemplo de señal modulada en amplitud en el dominio del tiempo

En la Ilustración 5Ilustración 11, es posible apreciar la forma de onda de la señal moduladora en la envolvente de la señal modulada. Este hecho no es característico de todos los métodos de modulación en amplitud, sino que únicamente será posible empleando la Ecuación 3 con un valor adecuado de índice de modulación (μ).

El índice de modulación representa la relación existente entre la amplitud de las señales moduladora y portadora. Para poder recuperar la moduladora a partir de la envolvente de la señal modulada, el valor máximo del índice será $\mu = 1$, caso el cual se encuentra representado en la Ilustración 11. Por encima de este valor, se dice que la señal experimenta sobremodulación.

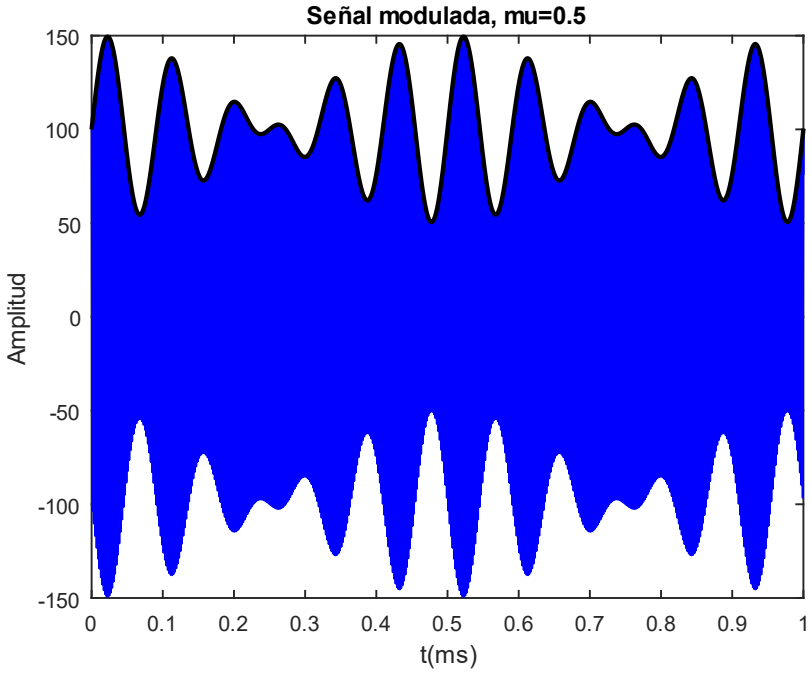


Ilustración 12. Ejemplo de señal modulada en amplitud con μ=0.5

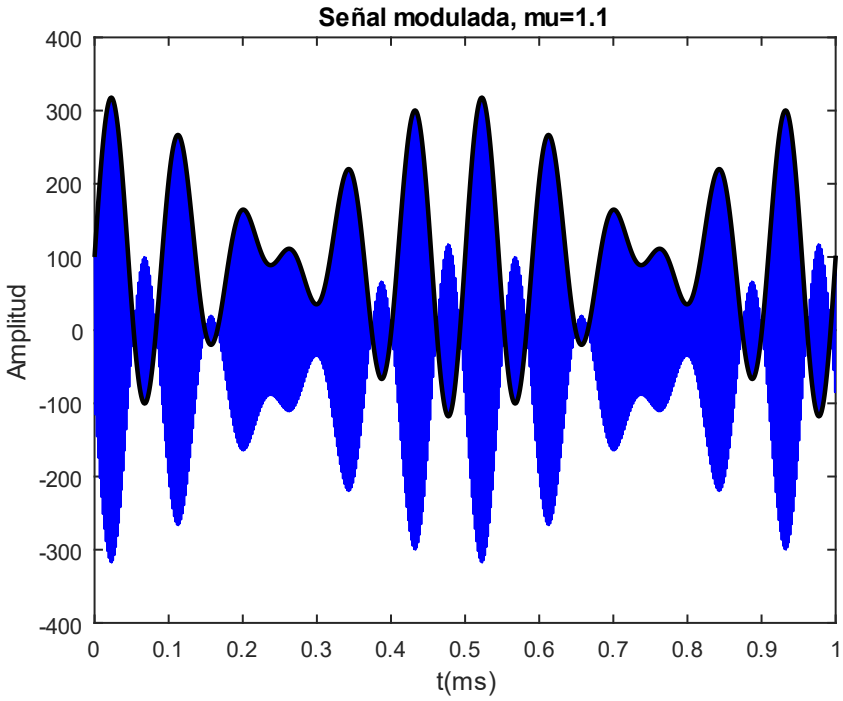


Ilustración 13. Ejemplo de señal modulada en amplitud con μ=1.1

La Ilustración 12 e Ilustración 13 ponen de manifiesto el efecto que la sobremodulación tiene sobre la envolvente de la señal modulada. En el caso para el que $\mu > 1$, la forma de la envolvente no coincide con la forma de onda de la señal moduladora.

Este hecho únicamente es crítico si el sistema está diseñado para recuperar la señal original a partir de un detector de envolvente, un circuito especialmente sencillo y económico de construir. Sin embargo, para el caso de estudio de este proyecto, no será necesario tener en cuenta este factor.

En el desarrollo del proyecto se emplearán las siguientes técnicas de modulación en amplitud: doble banda lateral con coseno, doble banda lateral simple, banda lateral única con filtro estándar y banda lateral única con filtro de Hilbert.

La manera de implementar cada una de estas técnicas, así como el efecto de las mismas sobre el espectro en frecuencia de la señal original, serán tratados en apartados posteriores.

Capítulo 3. ESTADO DE LA CUESTIÓN

3.1 PRIMERAS VULNERABILIDADES ENCONTRADAS

La introducción de los sistemas asistentes de voz en diferentes dispositivos de uso comercial va inevitablemente ligada a la búsqueda de vulnerabilidades, tanto intrínsecas a estos sistemas, como aquellas que pudieran aparecer en el dispositivo en cuestión a raíz de la incorporación de esta nueva herramienta.

Uno de los primeros asistentes de voz en el mercado fue Siri, desarrollado por la compañía Apple e incorporado por primera vez en uno de sus dispositivos en el año 2011, fecha de lanzamiento del Iphone 4s, sustentado por el sistema operativo IOS 5 (Allen, 2021)

Tres años más tarde, se encontraban las primeras vulnerabilidades relacionadas con Siri que afectaban a dispositivos equipados con versiones de IOS superiores a la 7.0. Por primera vez, se demostraba posible burlar los mecanismos de seguridad encargados del control de acceso al teléfono (el más extendido, el código PIN) para realizar operaciones que vulneran la privacidad del usuario, como la posibilidad de acceder a los contactos del teléfono, hacer llamadas y enviar e-mails. En este caso, una determinada secuencia de comandos, seguida de una reinserción de la tarjeta SIM, era suficiente para poder instruir al dispositivo, por medio de la voz, para ejecutar acciones que debieran haber requerido de autenticación previa (González, 2014).

Los diferentes dispositivos también se pueden ver afectados por herramientas *software* desarrolladas por terceros. El trabajo de Diao y otros (2014) pone de manifiesto una vulnerabilidad propia del sistema de búsqueda por comandos de voz de Google, *Google Voice Search,* que podía llegar a comprometer la privacidad de *smartphones* con sistema operativo Android. La técnica desarrollada por este grupo de investigadores consistía en emplear una aplicación desarrollada para Android y aparentemente segura, al no demandar al usuario ningún tipo de permiso de acceso a elementos sensibles del teléfono, como serían

la cámara o el micrófono. Esta aplicación era capaz de invocar a *Google Voice Search* y reproducir comandos de voz en segundo plano, los cuales eran reconocidos y ejecutados por la aplicación, esquivando de esta manera la jerarquía del sistema de permisos del dispositivo.

Podemos ver cómo, en ambos casos, se exponen vulnerabilidades en sistemas de reconocimiento por voz que se traducen en amenazas sobre la privacidad del dispositivo que los aloja. No obstante, debemos tener en cuenta que el nivel de riesgo de estas amenazas no es extremo, puesto que, además de ser subsanadas en parches de actualización de los correspondientes sistemas operativos de cada dispositivo, comparten dos factores principales que las caracterizan: en primer lugar, el atacante necesita acceder al dispositivo atacado (en el caso de Siri, se necesita poder extraer y reinsertar la tarjeta SIM del dispositivo) o forzar al usuario a realizar algún tipo de acción sobre el mismo (en el caso de GVS[7], se necesita descargar la aplicación que actuará de forma maliciosa); en segundo lugar, el usuario puede ser consciente del ataque mientras se está produciendo e interrumpirlo.

El nivel de riesgo de estas amenazas se elevaría exponencialmente si, tal como se propone en modelos de ataque posteriores, el atacante no necesitara manipular el dispositivo objetivo de ninguna manera o el sistema de ataque no requiriese ninguna acción perceptible por el usuario.

3.2 ONDAS ELECTROMAGNÉTICAS COMO VECTOR DE ATAQUE

De acuerdo con lo previamente expuesto, una forma de elevar el nivel de riesgo de la amenaza es diseñar el ataque de tal manera que sea imperceptible para cualquier persona próxima al dispositivo atacado. Uno de los principales vectores de ataque cuya viabilidad ha sido llevada a estudio es el empleo de ondas electromagnéticas. A lo largo de las últimas dos décadas, se han realizado múltiples experimentos que han permitido conocer con cierto nivel de profundidad el efecto que la radiación electromagnética tiene sobre los componentes de

[7] Acrónimo del sistema asistente de voz "*Google Voice Search*"

un sistema electrónico y cómo ese efecto se traduce en modificaciones del comportamiento original del sistema.

En el año 2015, Kasmi y Esteves encontraron una forma de conectar las características de las ondas electromagnéticas con las vulnerabilidades intrínsecas de los sistemas de reconocimiento por voz, desarrollando un método de ataque que incrementaba exponencialmente el riesgo de la amenaza. Su trabajo describe un método de inyección de comandos de voz no audibles lanzados contra programas de reconocimiento de voz instalados en dispositivos móviles tipo *smartphone*.

Este método aprovecha el hecho de que las primeras generaciones de dispositivos *smartphone* presentaban frecuentemente la capacidad de sintonizar frecuencias de radio y, en su mayoría, se valían de instrumentos periféricos, generalmente auriculares, para que actuasen como antena necesaria para captar la banda de frecuencias adecuada. Si además el set de auriculares empleado constaba de un micrófono incorporado, en el momento en el que se conectaba a un *smartphone* actuaba simultáneamente como antena FM y como interfaz de entrada de audio.

El modelo de ataque propuesto se basa en emitir comandos de voz modulados en amplitud sobre la banda de frecuencias de radio FM (80-108 MHz) de manera que fueran primero demodulados por los auriculares actuando como antena FM y, seguidamente, interpretados por el sistema de reconocimiento por voz que estuviese tratando los propios auriculares como punto *input* de señales de audio.

Al probar la efectividad de este modelo de ataque, se ponía de manifiesto un nuevo paradigma de amenaza en el cual el usuario afectado ya no es consciente del ataque, pues una de las características de las ondas electromagnéticas de alta frecuencia es que son imperceptibles para cualquier ser humano.

Como contraparte tenemos:

- El grupo de dispositivos que pueden ser afectados se limita a dispositivos con capacidad de sintonizar radio FM y con auriculares conectados actuando a modo de antena.
- Se requiere un equipamiento muy específico para poder llevar a cabo el ataque: un sistema emisor de radiofrecuencia en la banda FM y un amplificador con ganancia superior a los 50 W. A mayor potencia efectiva del transmisor, mayor será el coste propio del equipo necesario para asegurar la efectividad del modelo de ataque.

3.3 COMANDOS DE VOZ OCULTOS EN SEÑALES AUDIBLES

El empleo de ondas electromagnéticas de alta frecuencia no ha sido la única aproximación al problema de elaborar un modelo de ataque que resulte imperceptible para el usuario atacado. La capacidad de un usuario de percibir comandos de voz que traten de aprovechar el sistema de control de su dispositivo para ejecutar alguna operación maliciosa no depende únicamente de que el usuario pueda oír dichos comandos, sino de que pueda entender que las muestras de audio escuchadas suponen efectivamente una amenaza a su privacidad. En otras palabras, si una persona pudiese escuchar un comando de voz emitido directamente hacia su dispositivo, pero no entendiese su contenido o funcionalidad, no podría determinar si está siendo objetivo de algún tipo de ataque y, consecuentemente, es improbable que tomase algún tipo de medida defensiva al respecto. Esta situación podría derivar en un modelo de ataque de elevado riesgo.

Tomando como punto de partida la idea anteriormente expuesta, Vaidya y otros (2015) evaluaron la viabilidad de introducir comandos de voz “ocultos” en señales de voz audibles pero ininteligibles para un ser humano. Conceptualmente, su trabajo se sustenta en la hipótesis de que los elementos de un fragmento de audio que le confieren la cualidad de ser entendible por una persona no son los mismos elementos que un algoritmo de reconocimiento de voz utiliza para inferir si dicho fragmento de audio representa uno de los comandos que el sistema entiende como válidos.

Según este estudio, la gran mayoría de los sistemas de reconocimiento de voz utilizan un modelo matemático para representar aquellas muestras de audio que son captadas a la entrada del mismo. En principio, los investigadores tratan a estos sistemas como una "caja negra", de manera que se desconoce cómo se construye el modelo en cuestión. No obstante, se teoriza que un elemento clave en la construcción de este modelo son los MFCC, los cuales constituyen una representación de la muestra de audio en el espectro de frecuencias que trata de aproximarse a la respuesta ofrecida por el oído humano.

Partiendo de estos supuestos, consiguen desarrollar un mecanismo de transformación de fragmentos de audio, denominado *Audio Mangling*, que consiste en modificar una señal acústica con un determinado comando de voz (mediante reajuste de valores de sus MFCC) de tal manera que la señal alterada sigue siendo interpretada como un comando de voz por una herramienta software de reconocimiento de voz, pero pierde la cualidad de ser inteligible.

Carlini y otros (2016) extienden el alcance de los resultados del trabajo anterior al probar la efectividad del modelo de ataque descrito en un mayor número de escenarios plausibles y empleando versiones actualizadas de los sistemas de reconcomiendo de voz. Además, valoraron la posibilidad de un modelo de ataque de "caja blanca", donde el atacante conoce con precisión las operaciones del algoritmo de modelaje de voz empleado por el sistema objetivo, dando lugar a un escenario en el que la efectividad del ataque se ve incrementada.

El éxito de estos experimentos pone de manifiesto la viabilidad de modelos de ataque basados en el uso de comandos maliciosos incluidos dentro de señales de audio ininteligibles. Las implicaciones de los resultados obtenidos ponen sobre aviso de una amenaza para la privacidad e integridad de dispositivos con sistemas asistentes de voz instalados.

El nivel de riesgo de esta amenaza será directamente proporcional a la magnitud de las ventanas de tiempo en las que el dispositivo atacado sea susceptible de recibir instrucciones por comandos de voz. Algunos de estos instrumentos cuentan con un mecanismo de activación basado en un comando iniciador, el cual debe ser interpretado por el

procedimiento software con el objetivo de habilitar el uso del resto de comandos de voz. En estos casos, la probabilidad de éxito del ataque disminuiría, pues sería necesario efectuar dos ataques consecutivos: uno para el comando de activación y otro para el comando con la instrucción maliciosa que se desea ejecutar en el dispositivo atacado. Serán, por tanto, especialmente vulnerables a esta amenaza aquellos artilugios cuyo sistema de reconocimiento de voz se encuentre siempre activo, es decir, que procese cualquier señal acústica que sea captada por el punto de entrada de audio del dispositivo.

3.4 EMPLEO DE SEÑALES NO AUDIBLES: BACKDOOR ATTACKS

El empleo de ondas electromagnéticas como vector de un ataque para tratar de manipular la respuesta de un sistema asistente de voz ha dado lugar a diferentes estudios que tratan de complementar el trabajo de Mukhopadhyay y otros (2015).

Uno de las principales limitaciones del modelo de ataque propuesto por este grupo de investigadores es la elevada frecuencia de las ondas portadoras de la información que se desea hacer llegar al dispositivo objetivo. El propósito de emplear valores de frecuencia de gran magnitud es trabajar por encima del rango de frecuencias audible, de manera que la señal portadora de los comandos maliciosos que constituyen el ataque resulte imperceptible para cualquier ser humano. No obstante, la máxima frecuencia audible por un ser humano adulto promedio se estima próxima a los 20 kHz (Purves, y otros, 2012), de manera que si consideramos el rango de frecuencias a emplear como el rango FM (80-108 MHz), la frecuencia de las señales portadoras utilizadas se sitúa tres órdenes de magnitud por encima del valor umbral superior del espectro audible.

El valor de frecuencia de las ondas electromagnéticas emitidas es un parámetro clave con influencia directa sobre la facilidad de realización del ataque, pues a mayor frecuencia, mayor ganancia es necesaria en el transmisor para asegurar que la onda alcanza el receptor con un nivel de potencia superior al umbral aceptado. Transmisores de mayor ganancia implican a su vez un mayor coste, de manera que reducir la frecuencia de las señales

portadoras empleadas en el ataque incrementaría su viabilidad y, por tanto, el nivel de riesgo del mismo.

De acuerdo con esta línea de mejora, Roy y otros (2017) propusieron un modelo de ataque tipo *BackDoor*, por el cual es posible generar tonos de frecuencias dentro del rango audible a partir de señales con frecuencias situadas fuera de ese rango. Las frecuencias de las señales portadoras empleadas en este caso se sitúan dentro del intervalo 40-50 kHz, por lo que se encuentran notablemente más próximas al valor umbral estimado de 20 kHz, dando como resultado un ataque técnica y económicamente más viable que el propuesto por Mukhopadhyay y otros (2015).

La vulnerabilidad explotada por estos investigadores reside en la característica no lineal presente en la respuesta en frecuencia manifestada por los micrófonos que actúan como entrada de audio para la mayoría de dispositivos electrónicos. La hipótesis que da pie a este estudio es que el comportamiento no lineal de los micrófonos puede llegar a crear réplicas de las señales recibidas en frecuencias distintas de las originales, de tal manera que, si una de estas frecuencias se sitúa por debajo del umbral de 20 kHz, se estarían creando señales audibles a partir de señales no audibles.

Los resultados muestran que los micrófonos de los dispositivos empleados en la fase de experimentación (dispositivos tipo *smartphone*) exhiben una característica no lineal tal como se teorizaba, poniendo de manifiesto la viabilidad de un modelo de ataque por ondas ultrasónicas de frecuencias situadas en el mismo orden de magnitud que el umbral de audición humano.

El principal reto que encontraron los investigadores para desarrollar este modelo de ataque es que los componentes del transmisor también pueden presentar una respuesta no lineal similar a la del receptor. Si esto ocurriese, no se podría lograr el efecto deseado en la señal recibida puesto que, al ser emitida, ya habría experimentado modificaciones que hubieran alterado la forma de señal original.

Para salvar este inconveniente, los investigadores proponen combinar un tipo de modulación FM con la emisión de una segunda señal para conseguir el efecto deseado en recepción. En este caso, la viabilidad del ataque se vería reducida por la necesidad de usar, no uno, sino dos transmisores independientes para emitir señales de ataque hacia un objetivo.

3.5 EMPLEO DE SEÑALES NO AUDIBLES: DOLPHIN ATTACKS

Una de las principales aportaciones al estudio de vulnerabilidades, en lo que a sistemas de reconocimiento de voz se refiere, llega de la mano del artículo publicado por Zhang y otros (2017), el cual documenta el diseño y posterior prueba de un sistema de ataque contra dispositivos controlados por voz, basado en la generación de señales no audibles para el ser humano pero que, sin embargo, son interpretadas por un determinado grupo de asistentes de voz, llegando estos a ejecutar acciones en consecuencia.

El experimento se basa en la misma hipótesis contemplada anteriormente por Roy et al para los dispositivos móviles, aunque extrapolada a cualquier dispositivo que haga empleo de un asistente de voz. Según la misma, estos dispositivos emplean para la captación de sonido, micrófonos con una respuesta en frecuencia no lineal. A partir de esta hipótesis deriva la idea de aplicar una modulación en amplitud a una señal de voz utilizando una portadora de frecuencia superior a 20 kHz y aprovechar la supuesta característica no lineal de los micrófonos para realizar una "demodulación indirecta de la señal". Esto es, recuperar la señal de voz original dentro del circuito interno del dispositivo antes de atravesar el filtro paso-bajo, presente en todos los equipos de estas características.

Este filtro atenúa significativamente todos los componentes de frecuencia superiores a 20 kHz antes de digitalizar las señales y enviarlas al módulo en el que se aplican se los pertinentes algoritmos de reconocimiento de voz sobre la señal. Por ello, es esencial que la señal experimente este proceso de "demodulación indirecta" a partir de los componentes situados antes del filtro en el sistema de captación de audio del dispositivo.

Con esto, se conseguiría hacer llegar al subsistema de reconocimiento de discurso del dispositivo una señal de voz con un comando determinado, sin que esta haya podido ser escuchada por ningún ser humano situado en las proximidades del dispositivo atacado.

El experimento se llevó a cabo con los siguientes medios:

- Las señales de voz con los comandos a ejecutar son generadas empleando un software de conversión de texto a voz, instalado en un dispositivo móvil tipo *smartphone*. Se prueban, tanto herramientas software de diferentes distribuidores, como diferentes comandos de voz.
- Cada señal de voz generada se modula, aplicando una modulación tipo AM o de amplitud, empleando un dispositivo generador de señales. Los parámetros característicos de la modulación, como serían el índice de modulación y la frecuencia central de la portadora empleada, son elegidos en base a resultados experimentales. Se registran aquellos valores de cada parámetro que ofrecen un mejor resultado en términos de efectividad del ataque. Se pone como condición, sin embargo, que el menor componente en frecuencia presentado por la señal modulada debe encontrarse por encima del umbral de audición humano.
- Para la emisión de la señal modulada hacia el dispositivo objetivo del ataque se hace uso de un altavoz ultrasónico, con el objetivo evitar que la respuesta en frecuencia del mismo distorsione frecuencias superiores a 20 kHz.
- Por último, este sistema de ataque se aplica sobre un grupo de dispositivos con algún tipo de software de control por voz instalado. En total, se hace uso de 17 dispositivos diferentes, manufacturados por 8 fabricantes distintos.

Los resultados del experimento demostraron la efectividad del ataque diseñado sobre todos los dispositivos estudiados. También se determinaron los parámetros de la modulación con los que se conseguía una mayor efectividad en el ataque, así como la máxima distancia entre altavoz y dispositivo a partir de la cual el ataque dejaba de ser efectivo.

3.6 EMPLEO DE SEÑALES NO AUDIBLES: SURFING ATTACKS

La última aproximación que encontramos al problema de desarrollar un modelo de ataque basado en señales no audibles que logren inyectar comandos en sistemas de reconocimiento de voz se expone en el trabajo publicado por Yan y otros (2020).

Los resultados de esta investigación ponen de manifiesto la viabilidad de otro tipo de ataque no audible sobre dispositivos de control por voz: *SurfingAttack.* Este nuevo modelo de ataque amplía la limitación en la distancia efectiva que presentaba el modelo *Empleo* de señales no audibles: dolphin attacks, utilizando como medio de transmisión para las señales ultrasónicas una superficie sólida en lugar del aire.

Aprovechando las características de propagación de ondas electromagnéticas por un medio sólido, se define un modelo de ataque con diferentes ventajas respecto a los anteriormente propuestos: mayor rango efectivo de ataque, menor consumo de potencia por el transmisor y supresión de la necesidad de alinear físicamente emisor y receptor para una correcta recepción de las señales.

A partir de los resultados expuestos en este trabajo de investigación, se dibuja un escenario plausible en el que el dispositivo atacado y el dispositivo transmisor del atacante se encuentran simultáneamente en contacto con una superficie que permite la propagación de las señales de ataque.

Los investigadores resaltan dos condiciones esenciales para asegurar la viabilidad de este ataque:

- Las señales ultrasónicas propagadas a través de la superficie deben alcanzar el micrófono que el dispositivo atacado esté empleando como sensor de entrada de audio.

La propagación de ondas a través de materiales sólidos tiene asociada el fenómeno de distorsión acústica, por el cual la onda se separa en sus diferentes componentes de frecuencia al ser propagada por el material (diferentes frecuencias se propagan a diferente velocidad).

Para salvar este efecto, hay que elegir una forma de onda y un modo de propagación adecuado.

- No es necesario que el micrófono se encuentre en contacto directo con la superficie, aunque es esencial que la señal llegue con suficiente energía como para activar la característica no lineal del mismo.

Este trabajo demuestra que es posible aprovechar la característica no lineal del micrófono del receptor (que sirve como base para modelos de ataque descritos anteriormente) a partir de señales propagadas por un medio sólido. Para lograrlo, los investigadores se valen de un transductor piezoeléctrico para inducir vibraciones en el medio de propagación.

Capítulo 4. DEFINICIÓN DEL TRABAJO

4.1 JUSTIFICACIÓN

El **¡Error! No se encuentra el origen de la referencia.** expone un escenario con diferentes a rtículos de investigación centrados en un campo de estudio común: modelos de ataque contra dispositivos asistentes de voz. Cada uno de estos estudios toma como referencia el trabajo anterior, tratando de mejorar el modelo de ataque propuesto para aumentar el grado de concienciación y conocimiento de la industria tecnológica sobre amenazas que pueden llegar a poner en jaque la seguridad de los usuarios de estos dispositivos.

Uno de los aspectos de mayor relevancia a tener en cuenta a la hora de caracterizar un modelo de ataque es su nivel de riesgo. Este término hace referencia a un valor de medida que trata de reflejar con objetividad la probabilidad de que dicha amenaza llegue a materializarse. Asimismo, dicha probabilidad se encuentra estrechamente ligada a dos factores: la efectividad del ataque y la facilidad de realización del mismo.

4.1.1 FACTORES QUE INCREMENTARÍAN EL RIESGO DE ATAQUE

Debido a esto, un estudio que trabaje con hipótesis para incrementar el nivel de riesgo de un ataque no tendría por qué tratar únicamente con factores que pudieran incrementar el porcentaje de efectividad del mismo, sino que también podría considerar modificaciones en los paradigmas anteriores que hiciesen que el modelo descrito fuese más sencillo de poner en práctica por parte de un supuesto atacante.

Existen varios factores que podrían satisfacer esta última premisa: el uso de componentes de menor coste y/o más sencillos de adquirir en un entorno comercial, reducción de la complejidad de las tareas a realizar por el atacante o existencia de documentación anterior que pudiese ser reutilizada por el atacante para ahorrar tiempo en este proceso, entre otros.

Este proyecto se encuentra dedicado a estudiar un modelo de ataque basado en los paradigmas expuestos en el **¡Error! No se encuentra el origen de la referencia.**tomando c omo especial referencia el modelo de *Empleo* de señales no audibles: dolphin attacks y tratando de realizar una serie de adiciones que pudiesen aumentar el nivel de riesgo del mismo por efecto de los factores anteriormente enunciados. En los apartados subsiguientes se presentan las principales modificaciones a valorar.

4.1.1.1 Uso de dispositivos de última generación

Desde la publicación del documento expositivo del modelo *Empleo* de señales no audibles: dolphin attacks en el año 2017, no se ha realizado un estudio para valorar si los resultados obtenidos son extrapolables a dispositivos cuya fecha de lanzamiento es posterior a la realización del proyecto referenciado.

La industria de los dispositivos asistentes de voz no es ajena al ritmo del mercado en el que se mueven los productos tecnológicos de uso extendido, como los teléfonos tipo *smartphone*. Antes bien, debido que cierto número de asistentes actualmente se pueden encontrar integrados como herramientas *software* dentro de este tipo de dispositivos, se podría considerar que actualmente comparten cierta cuota de mercado.

Es bien sabido que las empresas dedicadas a manufacturar y comercializar este tipo de productos tienden a desarrollar nuevos prototipos con periodicidad anual, de manera que, desde el año 2017, son muchos los nuevos modelos de dispositivos que han visto la luz, acompañados en su mayoría de versiones actualizadas de sistemas de reconocimiento de voz ya conocidos.

Además, a consecuencia del elevado ritmo de consumo de los usuarios, no es descabellado suponer que los dispositivos de última generación aglomeran una parte importante del total de dispositivos en uso.

El hecho de estudiar si el modelo de ataque propuesto con anterioridad sigue siendo viable para las versiones más recientes de estos sistemas y sus correspondientes elementos *software* puede aportar información útil sobre el nivel de riesgo actual.

4.1.1.2 Aplicación de técnicas de modulación digital

El uso de herramientas *software* aplicadas al proceso de modulación de señal que constituye uno de los pasos necesarios del modelo de ataque puede tener implicaciones importantes sobre la viabilidad del mismo.

Por un lado, dada la naturaleza digital de estas herramientas, habría que considerar las ventajas de trabajar en este dominio con respecto del dominio analógico. Entre estas, cabría destacar la mayor flexibilidad y capacidad de modificación que tenemos sobre las señales digitales. Además, muchas de las aplicaciones que permiten realizar las operaciones de tratamiento de señal requeridas para llevar a cabo la operación de modulación pueden obtenerse de manera gratuita. Estos factores podrían suponer un incremento de la facilidad de realización del ataque.

Por otro lado, no se pueden obviar los inconvenientes intrínsecos al uso de este tipo de técnicas. El hecho de transformar la señal analógica al dominio digital y viceversa puede conllevar una pérdida de información en la señal que ocasione una disminución de la efectividad del ataque.

Como consecuencia, el estudio de la viabilidad de emplear técnicas de modulación digital podría aportar argumentos consistentes para caracterizar el nivel de riesgo del ataque.

4.1.1.3 Empleo de sistemas emisores de audio de carácter comercial

El paradigma de ataque mediante señales no audibles necesita, como su propio nombre indica, de la emisión de señales cuyos componentes en frecuencia se sitúen por encima del umbral auditivo humano.

En trabajos anteriores se recurre a dispositivos emisores de audio especialmente preparados para la emisión de señales en este rango de frecuencias. Sin embargo, el empleo de estos sistemas tiene dos inconvenientes principales: su elevado coste (el dispositivo emisor empleado en el desarrollo del modelo de *Empleo* de señales no audibles: dolphin attacks tiene un precio de mercado actual de 450€ (Avisoft Bioacustics, s.f.)) y la dificultad asociada

a su adquisición, ya que son comercializados por un número reducido de distribuidores. Estos factores repercuten negativamente en la facilidad de realización del ataque.

El uso de altavoces de carácter comercial podría considerarse una alternativa que, de ser viable, aumentaría de seguro el nivel de riesgo del ataque, al tratarse de dispositivos más baratos y fácilmente adquiribles. No obstante, la respuesta en frecuencia de estos aparatos no asegura una característica lineal en el rango de frecuencias considerado por el modelo de ataque, por lo que se hace necesario un estudio de viabilidad antes de poder extraer cualquier conclusión.

El estudio de la viabilidad de estas variaciones puede servir para ofrecer una imagen actualizada del nivel de riesgo al que están sometidos los dispositivos asistentes de voz.

Cualquier modificación en el nivel de riesgo asociado a una determinada amenaza debe ser tenida en consideración, pues la materialización de la misma sobre un grupo de dispositivos concreto puede tener consecuencias catastróficas para la industria. Basta con tomar el ejemplo reciente de Intel[8] para entender la magnitud de las repercusiones negativas, principalmente a nivel de activos económicos y mala imagen en prensa, que un fallo de seguridad ocasiona en una empresa.

La importancia de evitar situaciones como las comentadas se considera la principal justificación sobre la que se sustenta la realización del proyecto.

4.2 OBJETIVOS

Este trabajo pretende proponer y estudiar un modelo de ataque sobre dispositivos asistentes de voz mediante el empleo de señales no audibles. Se tomarán como referencia estudios anteriores sobre modelos similares, tratando de incorporar una serie de modificaciones que

[8]En el año 2018, las acciones de la compañía Intel disminuyeron su valor de mercado en torno a un 6% tras la detección de una vulnerabilidad en sus procesadores (Quintanilla, 2018)

podrían incrementar el nivel de riesgo del mismo. Por ello, se han fijado los siguientes objetivos:

Diseño de un modelo de réplica

El primer objetivo que se pretende alcanzar con el proyecto es el desarrollo de un modelo que permita replicar el ataque de control por señales no audibles sobre un dispositivo asistente de voz.

Para el desarrollo de este modelo de ataque, se estudiarán diversos factores susceptibles de incrementar su nivel de riesgo, como el uso de técnicas de modulación digital o el empleo de dispositivos emisores de audio de carácter comercial.

Replicar el modelo de ataque DolphinAttack

Se pretende replicar, bajo condiciones controladas, el modelo de ataque diseñado y documentado por Zhang y otros en el año 2017 [apartado 3.5]. Este objetivo va ligado a un estudio de la viabilidad del ataque, para el cual se tendrá en cuenta la probabilidad de éxito del mismo, los factores que incrementan o decrementan su efectividad y el grado de accesibilidad al dispositivo atacado, si el ataque resulta ser satisfactorio.

Obtener información que permita caracterizar el funcionamiento de distintos asistentes de voz

Se pretende diseñar una batería de pruebas que permitan obtener más información acerca del funcionamiento de distintos dispositivos asistentes de voz. Los resultados obtenidos podrían ser empleados para predecir nuevas vulnerabilidades o proponer futuras líneas de estudio en relación a posibles medidas de defensa.

4.3 METODOLOGÍA Y PLANIFICACIÓN

Se ha divido el desarrollo del trabajo en cinco fases diferenciadas, expuestas en la Ilustración 14, orientadas a cumplir con todos los objetivos anteriormente expuestos:

Fase 0: Documentación previa

Esta fase comprende el proceso de recopilación y lectura de documentación relacionada con el marco de estudio del proyecto, previa a la realización del mismo. La fase se dará por concluida con la entrega del Anexo B, a mediados del mes de noviembre.

Fase 1: Diseño del modelo de ataque

En esta fase se procederá al diseño y preparación del experimento que consistirá en la replicación de un modelo de ataque *DolphinAttack* sobre un dispositivo controlado por voz.

Se realizarán diferentes pruebas para comprender mejor el funcionamiento, por, separado, de los diferentes elementos que serán necesarios para llevar a cabo el experimento. Esto es, un software de generación de señales de voz (por ejemplo), un programa para modular una señal de voz en amplitud (inicialmente, se ha considerado para su desarrollo el uso del entorno de programación Matlab) y un sistema de emisión de audio de uso comercial.

Fase 2: Obtención información que permita caracterizar el funcionamiento de distintos asistentes de voz

El propósito de esta fase es describir y realizar una batería de pruebas para obtener datos que permitan caracterizar el funcionamiento de diferentes dispositivos asistentes de voz. A la realización de cada prueba, se redactará un apartado de conclusiones para resumir los resultados obtenidos y explicar las posibles implicaciones de los mismos.

Fase 3: Prueba del modelo diseñado

El objetivo a completar en esta fase será la puesta en práctica del diseño efectuado en la fase anterior. Para ello, se emplearán los recursos mencionados anteriormente y se hará uso, además, de uno o varios dispositivos con un software de asistente de voz instalado que harán las veces de objetivo del ataque.

Se comprobará de forma experimental la efectividad del ataque bajo unas condiciones definidas, así como los parámetros que influyen en el aumento o decremento de la misma.

Fase 4: Elaboración del informe final

Finalmente, se procederá a la recopilación de toda la información, diseños y resultados obtenidos durante la realización del trabajo para llevar a cabo la elaboración del informe final del proyecto.

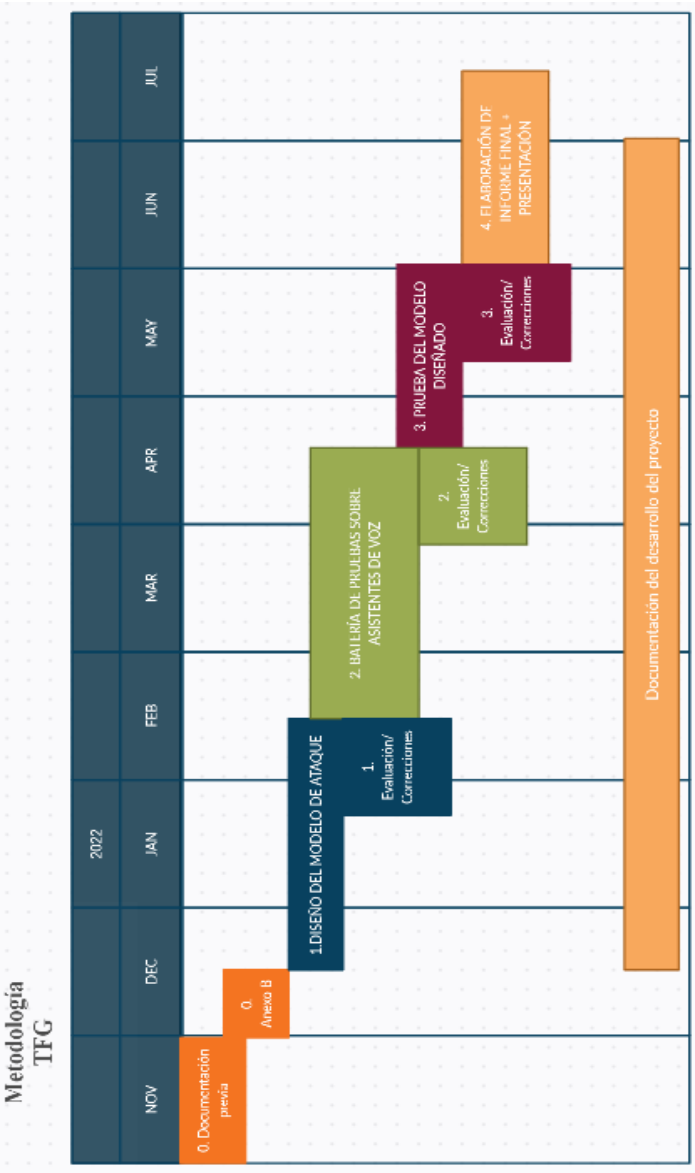


Ilustración 14. Planificación del proyecto

Capítulo 5. SISTEMA DESARROLLADO

5.1 HIPÓTESIS DE TRABAJO

El modelo de ataque a diseñar debe tener en cuenta que el trabajo se sustenta sobre la siguiente hipótesis:

La respuesta en frecuencia de algunos componentes electrónicos en el receptor presenta una característica no lineal para frecuencias superiores a 20 kHz.

El efecto de una respuesta no lineal sobre una señal modulada puede ocasionar que aparezcan réplicas de la señal en frecuencias distintas de la frecuencia de portadora. Si una de estas réplicas se sitúa en banda base y consigue atravesar los filtros en potencia y frecuencia del sistema del dispositivo objetivo, dicha señal puede ser interpretada como un comando normal, emitido en banda base.

De acuerdo con la información desarrollado en el apartado 2.1.1, el subsistema de captación de audio del dispositivo asistente de voz es muy probable que cuente con un filtro *antialiasing*[9] de frecuencia de corte inferior a 20 kHz. Esto implica que los resultados de una transformación no lineal de la señal modulada deberían ser efectivos en los componentes previos a este filtro: micrófono y amplificador.

Una manera sencilla de representar el efecto de una transformación no lineal en un componente electrónico es incluir un término cuadrático en la ecuación que define la respuesta del sistema. Este término puede añadirse sumando al término lineal que originalmente debería aparecer a la salida del componente en cuestión. De esta forma, si

[9] Filtro paso-bajo cuya función es limitar la frecuencia máxima de portadora para evitar el solapamiento espectral tras el proceso de muestreo

consideramos la señal modulada $y(t)$, a su paso por un componente en régimen de no linealidad la salida podría describirse como:

Ecuación 4 $$s(t) = A\,y(t) + By(t)^2$$

siendo A y B las ganancias para los términos lineal y cuadrático respectivamente.

En régimen de trabajo lineal, donde idealmente deben trabajar todos los componentes del sistema, se tiene que la relación entre los términos de ganancia reflejados en la Ecuación 4 es A >> B. De esta manera, el término cuadrático tiene un peso inapreciable en la salida.

No obstante, una vez se entra en la zona de trabajo no lineal, las magnitudes de estos dos términos pueden equipararse para determinadas frecuencias. Es entonces cuando se aprecia el efecto de la no linealidad sobre la señal.

Este efecto puede ocasionar que aparezcan aportaciones en frecuencia que no aparecían en la señal original. Se puede comprobar fácilmente si consideramos $y(t)$ como un tono básico. Si $y(t) = \cos(2\pi f t)$, el término cuadrático se podría escribir como :

$$B\,cos^2(2\pi f t) = B\,(1 + \cos(2\pi\,(2f)t)$$

En este caso, aparece un término en el doble de la frecuencia fundamental que originalmente tenía la señal $(2f)$. Los efectos son más evidentes si en lugar de un tono puro, la señal $y(t)$ es el resultado de modular en amplitud una señal periódica.

Tomando y(t) como el resultado de modular en amplitud un tono puro, empleando el método de doble banda lateral con coseno , se obtienen los siguientes resultados:

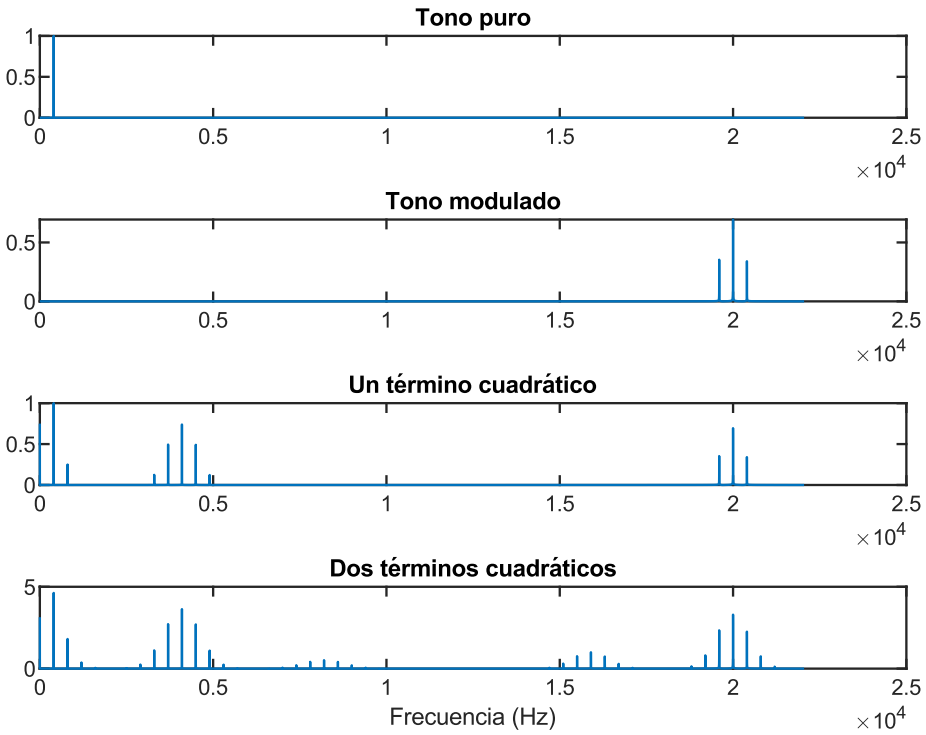


Ilustración 15. Efecto de una respuesta en frecuencia no lineal sobre un tono puro modulado en amplitud

Como se puede observar en la , la agregación de un término cuadrático genera aportaciones en frecuencia que no son propias de la señal modulada. Una de estas aportaciones puede resultar en una réplica de la señal original en banda base.

Cuanto mayor sea el número de componentes con respuesta en frecuencia no lineal que deba atravesar la señal, mayor será el número de nuevas frecuencias que se podrán registrar en la señal a la salida del sistema.

5.2 DISEÑO Y DESARROLLO

El sistema de ataque ha sido diseñado emulando el modelo empleado para reproducir un *DolphinAttack.* A continuación, se describen los subsistemas que componen la estructura del modelo de ataque desarrollado, tal como se muestran en la Ilustración 16.

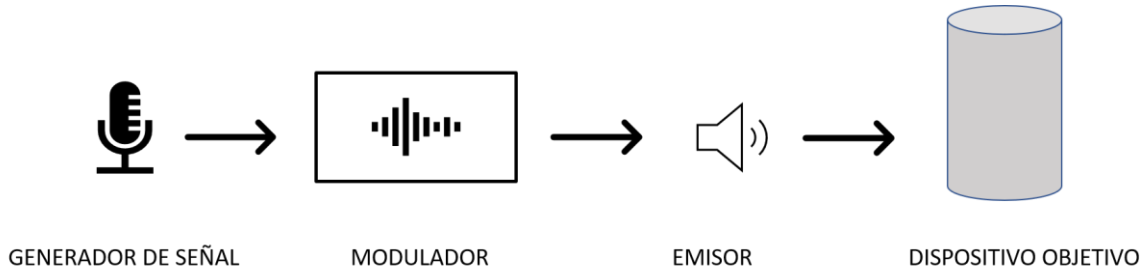


Ilustración 16. Esquema del modelo de ataque desarrollado

5.2.1 GENERADOR DE SEÑAL

El primer paso para generar un ataque sobre un dispositivo asistente de voz es generar la señal que contendrá el comando de voz que se quiere ejecutar en el dispositivo objetivo. Para generar esta señal de voz se contemplan dos opciones fundamentales: generación natural o generación artificial.

En el caso de generar la señal de voz de manera natural, se debe disponer de un sujeto que enuncie de manera clara y potente el comando de voz deseado. Además, será necesario un sistema de grabación que permita recoger la señal de voz con calidad suficiente para asegurar que sus parámetros básicos no se ven alterados y sigue siendo entendible por cualquier sistema de reconocimiento de voz.

En cuanto a la alternativa de generar el comando de voz de forma artificial, la opción más habitual es usar una herramienta software que permita convertir texto escrito en un fragmento de audio, imitando una voz humana. En la actualidad, el mercado de este tipo de herramientas es bastante amplio, existiendo opciones gratuitas.

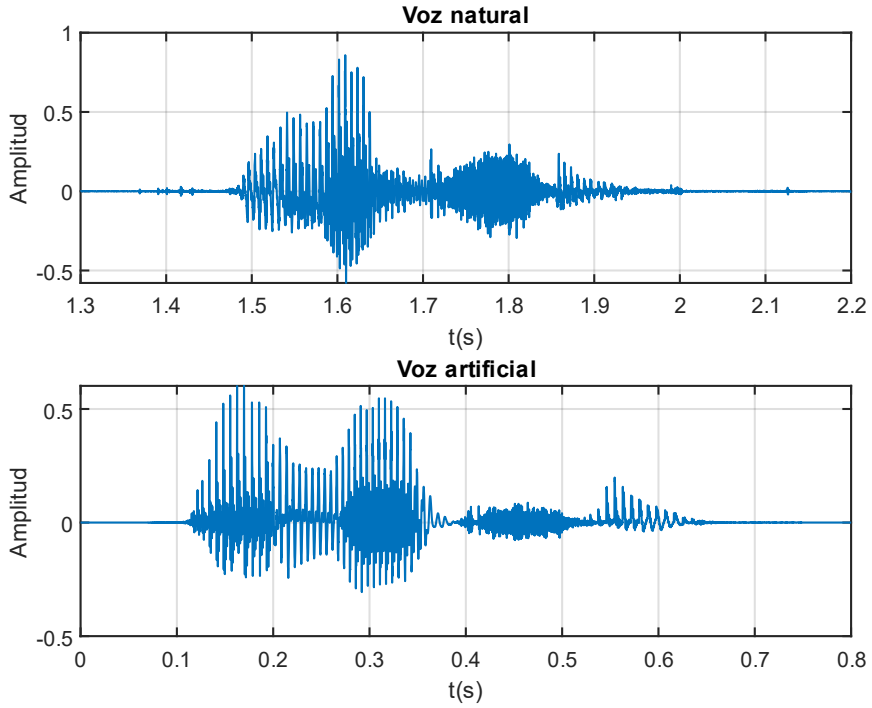


Ilustración 17. Comparación entre una señal de voz natural y una artificial en el dominio temporal

La comparación mostrada en la Ilustración 17 no pone de manifiesto ninguna diferencia significativa directamente relacionable con la efectividad de los mismos a la hora de ser reconocidos por un software especializado. Únicamente es remarcable que en la señal artificial es posible distinguir las diferentes sílabas de la palabra de manera más clara, tal como se puede apreciar en los intervalos de tiempo en los que hay una diferencia considerable de amplitud.

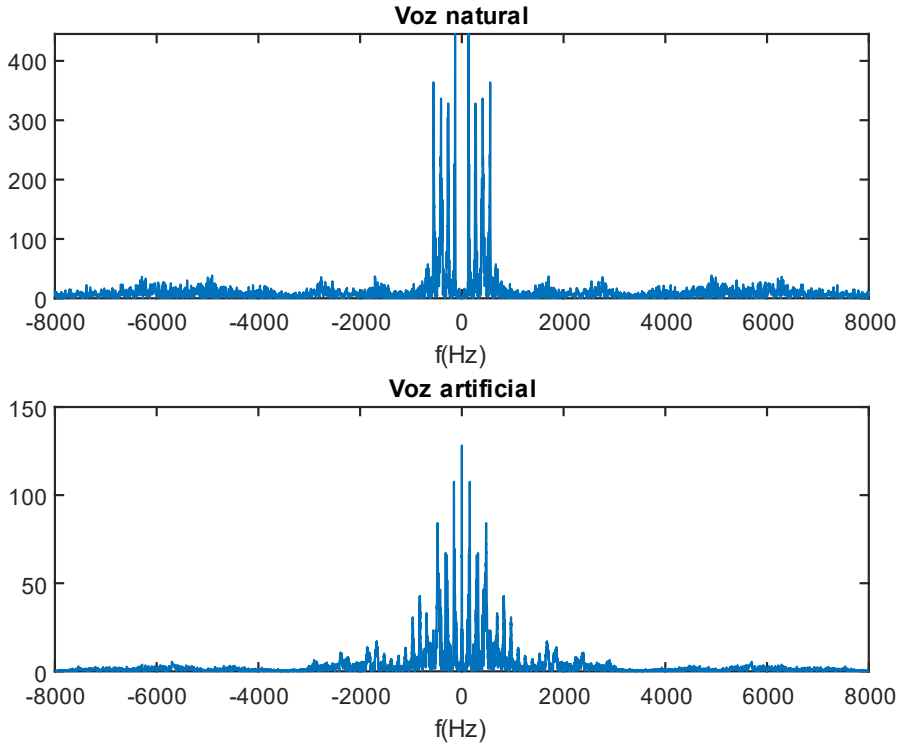


Ilustración 18. Comparación entre una señal de voz natural y una artificial en el dominio de la frecuencia

Una comparación de ambas señales en frecuencia tampoco permite dilucidar qué tipo de señal sería más apropiada para emplear en este modelo de ataque, puesto que el rango de frecuencias abarcado por ambas señales es aproximadamente el mismo. Es cierto que la señal de voz natural presenta un cierto nivel de ruido de alta frecuencia, producto del entorno y las condiciones de grabación, que no se muestra en la señal artificial, debido a la construcción sintética de la misma. Sin embargo, el empleo de un filtro paso-bajo sobre la señal de voz natural debería equiparar el nivel de ruido de ambas señales, de manera que este factor tampoco debería ser crítico para la efectividad del ataque.

El factor más importante que deberá tener cualquier señal que se quiera emplear como vector de ataque será su inteligibilidad. Tal como queda indicado en el apartado Capítulo 2. , toda señal captada por el dispositivo es modelada matemáticamente mediante algún tipo de

transformación, siendo uno de los métodos con mayor consenso de uso la extracción de los MFCC.

	c_1	c_2	3	c_4	c_5
Voz natural	-5.91	-21.8	2.04	1.37	0.78
Voz artificial	-374.37	-1000	0.97	-0.2	0.23

Tabla 1. Primeros 5 valores de los MFCC de una señal de voz natural y una artificial

Los MFCC tratan de ofrecer una transformación en frecuencia adaptada a la forma en la que el oído humano percibe las señales. Tras extraer los MFCC de la señal natural y la artificial (representados en la Tabla 1) y comparar los 5 primeros coeficientes con más peso en la transformación, se puede comprobar que existe una diferencia significativa entre los valores obtenidos para cada una de las señales. Este hecho lleva a deducir que ambas señales serán interpretadas de manera diferente por un mismo software de reconocimiento de discurso.

Ante la imposibilidad de determinar la señal idónea antes de realizar las pruebas planteadas, se ha decidido optar por el uso de señales de voz naturales y plantear el uso de señales de voz artificiales como un posible trabajo futuro.

5.2.2 MODULADOR DE SEÑAL

Uno de los objetivos de este proyecto es comprobar si la modulación digital es una posibilidad viable para reproducir un ataque inaudible sobre un dispositivo de control por voz (apartado 4.1.1.2). Para desarrollar una herramienta *software* con capacidad para reproducir técnicas de modulación en amplitud en el dominio digital, se ha usado el entorno de programación ofrecido por la herramienta Matlab.

El sistema modulador diseñado para este modelo de ataque tiene la capacidad de generar señales moduladas en amplitud mediante diferentes métodos, de manera que en el entorno

de pruebas se pueda comprobar la validez o medir el grado de idoneidad de cada uno de ellos.

Los diferentes métodos para generar una señal modulada de tipo AM difieren principalmente en la forma por la cual son implementados y en el aspecto del espectro de frecuencias de la señal modulada que se genera. A continuación, se procede a enumerar y explicar los diferentes métodos utilizados, partiendo de una señal cuyo espectro en frecuencia tiene el aspecto mostrado en la Ilustración 19 :

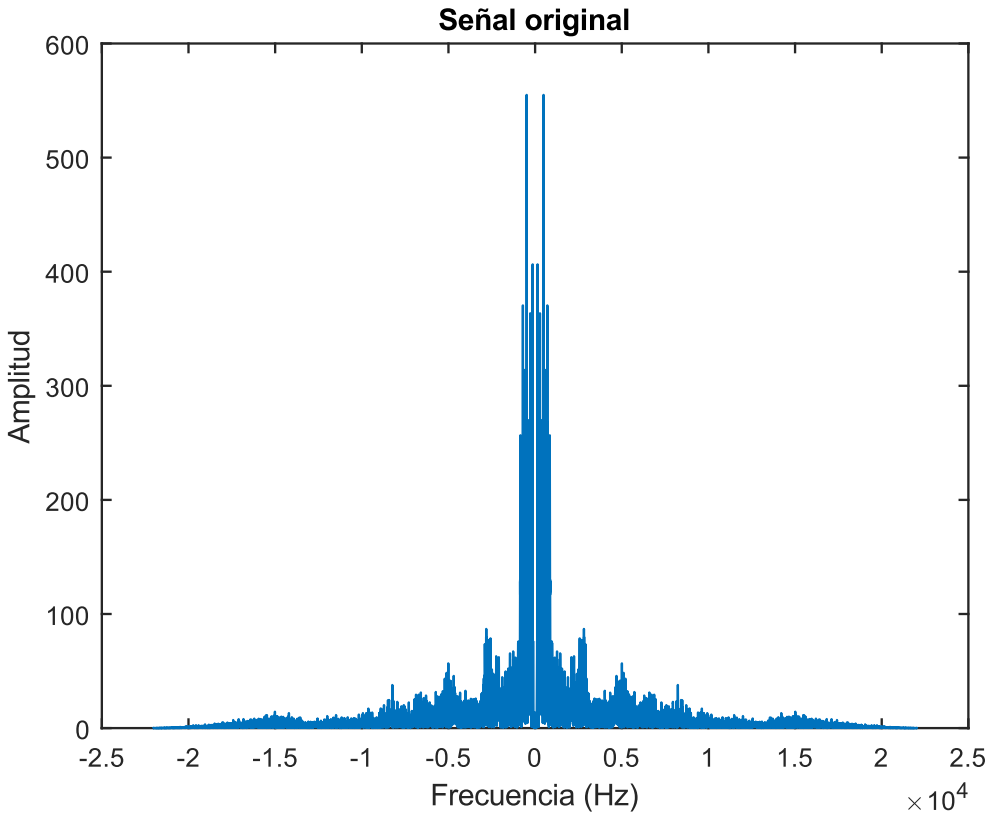


Ilustración 19. Módulo del espectro en frecuencia de una señal de voz en banda base

5.2.2.1 Doble banda lateral con coseno

Ecuación 5 $$y(t) = [1 + \mu\, m(t)]\ cos(2\pi f_w t) = \ cos(2\pi f_w t) + \mu\, m(t)\ cos(2\pi f_w t)$$

La obtención de la señal modulada con respecto del tiempo se puede modelar mediante la expresión matemática desarrollada en la Ecuación 5, siendo $m(t)$ la señal moduladora, μ el índice de modulación, f_w la frecuencia de la portadora e $y(t)$ la señal modulada. Este método ha sido implementado en un entorno digital mediante el siguiente fragmento de código:

```
%% Función para modular una señal en amplitud con método de Doble Banda Lateral
con coseno
%
%% Argumentos de entrada:
    % x_t = señal moduladora
    % mu = índice de modulación
    % Fw = frecuencia de portadora
    % t = vector de tiempos
    % A = ganancia de la modulación

function [y_t] = mod_am_cos(x_t,mu,Fw,t,A)

    port = cos(2*pi*Fw*t); % señal portadora
    y_t = A*(1 + mu*x_t).* port ; % señal modulada

end
```

Si se considera la notación $M(f), Y(f)$ como una representación de las transformadas de Fourier de $m(t), y(t)$ respectivamente, es posible obtener una representación matemática que muestra el efecto de la modulación sobre el módulo del espectro de la señal original (Ecuación 6)

Ecuación 6 $$Y(f) = \frac{1}{2}[\delta(f - f_w) + \delta(f + f_w)] + \frac{\mu}{2}[M(f - f_w) + M(f + f_w)]$$

Siendo δ la representación matemática de la señal Delta de Dirac.

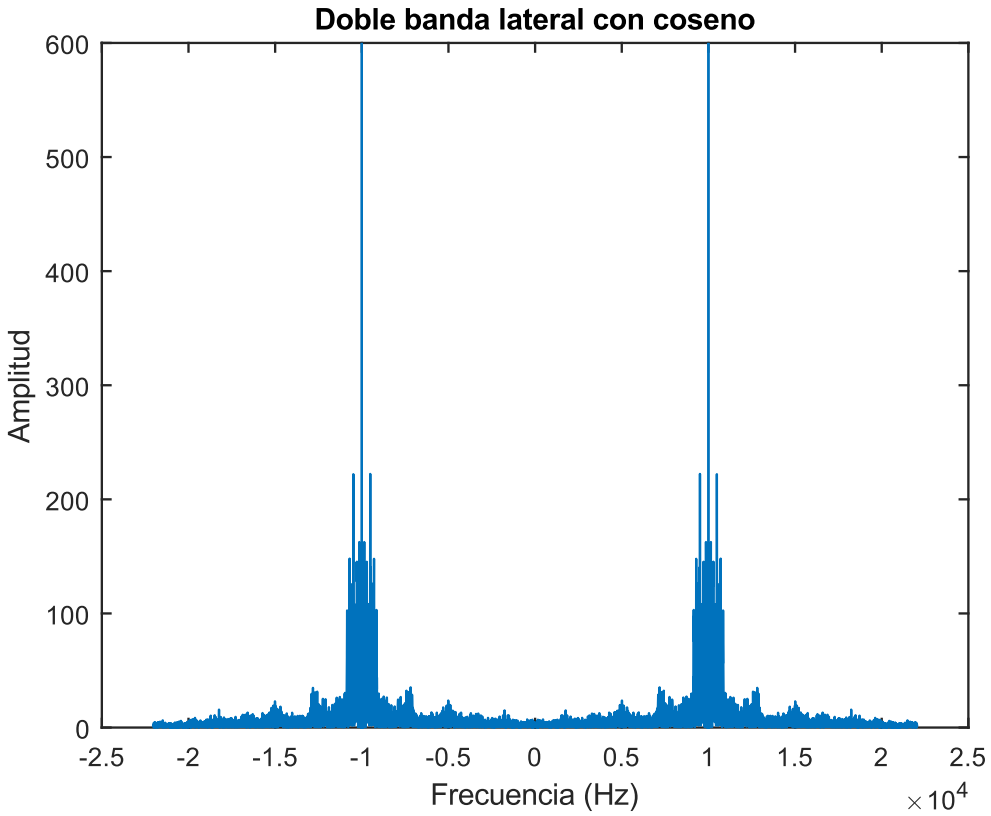


Ilustración 20. Módulo del espectro en frecuencia de una señal de voz modulada en doble banda lateral con coseno

Tal y como se muestra en la Ilustración 20, una de las particularidades de este método reside en que el espectro de la señal se encuentra duplicado con ambas bandas de frecuencia intactas y centradas en la frecuencia de la portadora. En este valor de frecuencia se pueden observar dos tonos de mayor energía que el resto de la señal, fruto de los términos $\delta(f - f_w) + \delta(f + f_w)$ que aparecen en la Ecuación 6. A su vez, estos términos son el resultado de aplicar la transformada de Fourier al término $\cos(2\pi f_w t)$.

La principal característica distintiva de este método es que el valor de μ, denominado como índice de modulación, permite controlar la relación de proporciones entre la amplitud de las señales moduladora y portadora. Un valor de $\mu < 1$ ocasiona que la forma de onda de la señal moduladora aparezca en la envolvente de la señal modulada, pudiendo esta ser recuperada mediante un sistema detector de envolvente.

La sencillez y bajo coste de implementación de este sistema en un dispositivo receptor hacen que este método de modulación en amplitud sea especialmente atractivo si se dispone de control sobre el diseño de emisor y receptor en la comunicación. Sin embargo, el modelo de ataque de este proyecto no considera ningún tipo de control sobre el aparato receptor, de manera que esta particularidad no es crítica para su diseño e implementación.

5.2.2.2 Doble banda lateral simple

Ecuación 7 $$y(t) = A\, m(t)\, cos(2\pi f_w t)$$

La obtención de la señal modulada con respecto del tiempo se puede modelar mediante la Ecuación 7, siendo $m(t)$ la señal moduladora, A la ganancia total de la modulación, f_w la frecuencia de la portadora e $y(t)$ la señal modulada.

La implementación en un entorno digital ha sido posible mediante el siguiente fragmento de código:

```
%% Función para modular una señal en amplitud con método de Doble Banda Lateral
simple
%
%% Argumentos de entrada:
    % x_t = señal moduladora
    % Fw = frecuencia de portadora
    % t = vector de tiempos
    % A = ganancia de la modulación

function [y_t] = mod_am_DBL_simple(x_t,Fw,t)

    y_t= A*x_t.*cos(2*pi*Fw*t); %Señal modulada

end
```

Reutilizando la notación empleada en la Ecuación 6, es posible obtener una representación matemática que muestra el efecto de la modulación sobre el módulo del espectro de la señal original (Ecuación 8)

Ecuación 8 $$Y(f) = \frac{A}{2}[M(f - f_w) + M(f + f_w)]$$

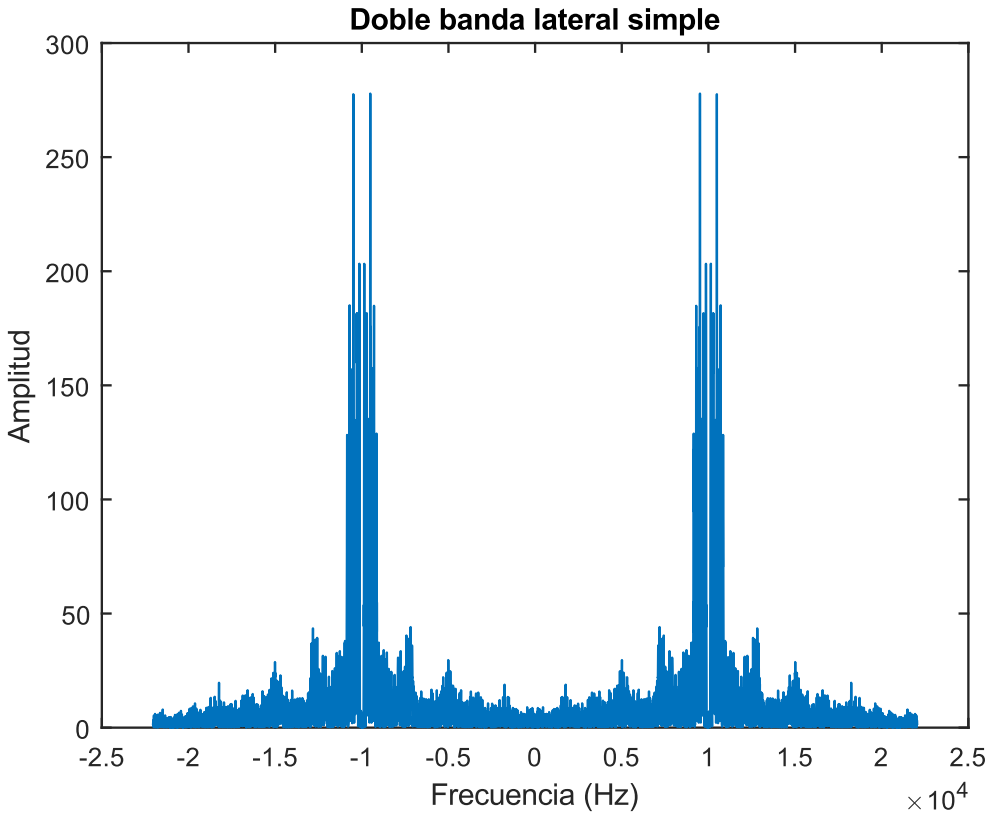


Ilustración 21. Módulo del espectro en frecuencia de una señal de voz modulada en doble banda lateral simple

Como se puede observar en , la forma del espectro de la señal modulada será idéntica al caso anterior, salvo por la ausencia de los tonos de alta potencia situados en los múltiplos positivos y negativos de las frecuencias de la portadora.

La energía total de la señal modulada es menor que en el caso de doble banda lateral con coseno y la señal original no podrá ser recuperada de forma sencilla mediante un sistema detector de envolvente.

5.2.2.3 Banda lateral única con filtro estándar

La modulación de banda lateral única reduce la energía total de la señal enviada y la redundancia de información en el canal mediante la eliminación de una de las bandas en las réplicas de la señal original. Para ello, se hace uso de algún tipo de filtro que actúe sobre las frecuencias que se desean atenuar.

Ecuación 9 $$y(t) = A\, m(t)\, cos(2\pi f_w t) * f(t)$$

La obtención de la señal modulada con respecto del tiempo se puede modelar mediante la Ecuación 9, siendo el carácter * la representación del operador "convolución" y $f(t)$ la expresión temporal del filtro empleado.

Reutilizando la notación empleada en la Ecuación 6, es posible obtener una representación matemática que muestra el efecto de la modulación sobre el módulo del espectro de la señal original (Ecuación 10), donde $H(f)$ represente el módulo del espectro en frecuencia del filtro empleado.

Ecuación 10 $$Y(f) = \frac{A}{2}[M(f - f_w) + M(f + f_w)]\, H(f)$$

En este caso, se está empleando un filtro directamente sobre la señal modulada en doble banda lateral, por lo que la implementación en código de este tipo de modulación requiere de un paso previo que reutilice la modulación en doble banda anteriormente expuesta.

En función de si se quiere conservar la banda de frecuencia superior o inferior de las réplicas de la señal, se empleará un tipo de filtro u otro. Será necesario un filtro paso-bajo para conservar la banda lateral superior y un filtro paso-alto para conservar la banda laterial inferior, siendo en ambos casos la frecuencia de corte del filtro igual a la frecuencia de portadora de la modulación ($f_w = f_c$)

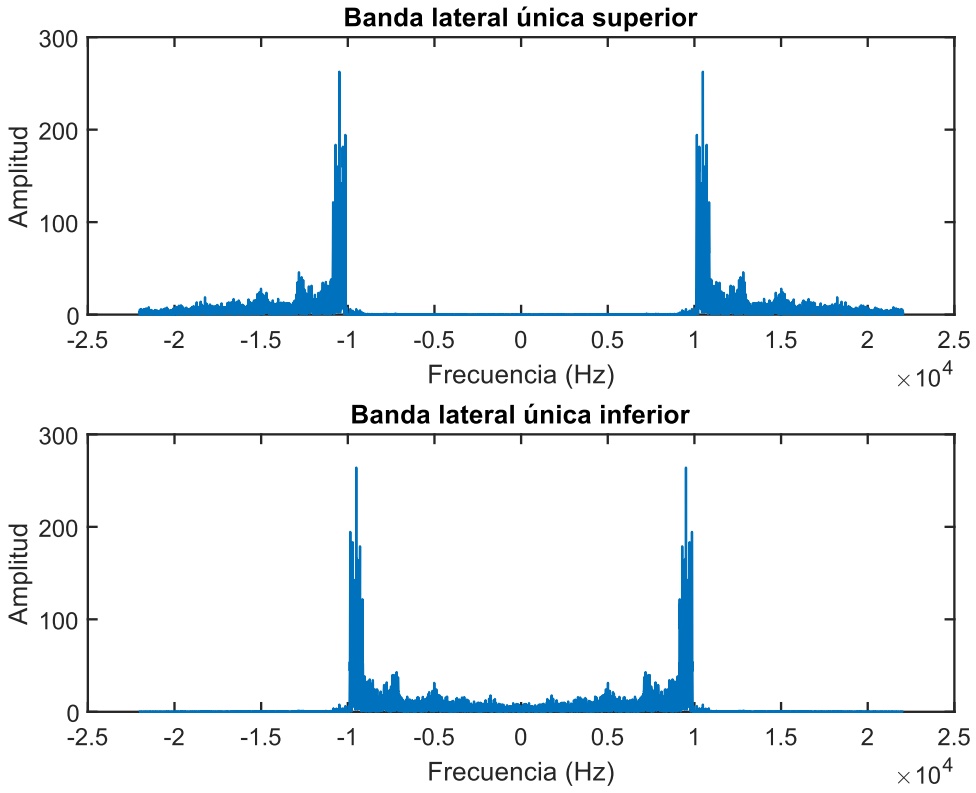


Ilustración 22. Módulo del espectro en frecuencia de una señal de voz modulada en doble banda lateral simple

La expresión matemática ideal para el espectro en frecuencia de sendos filtros sería:

$$H_{paso-bajo}(f) = G \quad \forall f < f_w$$

$$H_{paso-bajo}(f) = 0 \quad \forall f > f_w$$

$$H_{paso-alto}(f) = G \quad \forall f > f_w$$

$$H_{paso-alto}(f) = 0 \quad \forall f < f_w$$

En la Ilustración 22, se puede observar como las señales moduladas con este método se corresponden con réplicas del espectro de la señal original, aunque carentes de una de sus mitades. Como consecuencia, se reduce la energía total de la señal generada, pero se requiere mayor capacidad de computación, debido a que el volumen de operaciones a realizar es más alto que en los casos anteriores por el empleo de una fase de filtrado digital.

Este factor repercutiría negativamente si el procesado de las señales de ataque tuviera que ser realizado en tiempo real. Sin embargo, el modelo de ataque diseñado contempla la existencia de un lapso de tiempo indefinido entre el procesamiento digital de la señal de ataque y su posterior emisión, de manera que este hecho tampoco afecta de manera crítica a la efectividad del ataque.

Para la implementación de esta técnica de modulación en formato digital se ha desarrollado la siguiente función:

```
%% Función para modular una señal en amplitud con método de Banda Lateral Única
con filtro estándar
%
%% Argumentos de entrada:
    % x_t = señal moduladora
    % Fw = frecuencia de portadora
    % OPT = variable para elegir entre banda lateral superior o inferior
    % Fs = frecuencia de muestreo de la señal
    % t = vector de tiempos

function [y_t] = mod_am_BLU_coseno(x_t,Fw, OPT,Fs,t)

    y= x_t.*cos(2*pi*Fw*t); % Señal modulada en doble banda lateral

    Fpass = Fw-50; % Frecuencia de inicio de la banda de transción del filtro
(Hz)
    Fstop = Fw+50; % Frecuencia de fin de la banda de transción del filtro (Hz)
    Ap = 1;        % Rizado deseado en la banda de paso del filtro (en dB)
    Ast = 30;      % Atenuación deseada en la banda de atenuación del filtro (en
dB)

    % DISEÑO DE FILTRO FIR CON LAS CARACTERÍSTICAS ESPECIFICADAS

    if OPT == 0
        d = get_FIR_filter(Fpass,Fstop,Ap,Ast,Fs,0) % Banda lateral superior

    elseif OPT == 1

        d = get_FIR_filter(Fpass,Fstop,Ap,Ast,Fs,1) % Banda lateral inferior
```

```
    end

    %Aplicación del proceso de filtrado
    D = round(mean(grpdelay(d)));
    y_t= filter(d,[y;zeros(D,1)]);

    %Corrección del retardo introducido por el filtro
    y_t= y_t(D+1:end);

end
```

5.2.2.4 Banda lateral única con filtro Hilbert

Ecuación 11 $$y(t) = m(t)\,cos(2\pi f_w t) + [m(t) * hl(t)\,]\,sen(2\pi f_w t)$$

Este método ofrece un resultado idéntico al inmediatamente anterior, siendo la única diferencia entre ambos la forma de implementación. En este caso, la expresión matemática que representa la forma de obtener la señal modulada a partir de la moduladora se encuentra reflejada por la Ecuación 11, donde $hl(t)$ representa la expresión temporal para la transformada de Hilbert.

La implementación de este método de modulación en un entorno digital se consigue mediante el siguiente fragmento de código:

```
%% Función para modular una señal en amplitud con método de Banda Lateral Única
con filtro tipo Hilbert
%
%% Argumentos de entrada:
    % x_t = señal moduladora
    % Fw = frecuencia de portadora
    % fo = orden deseado para el filtro de Hilbert
    % Fs = frecuencia de muestreo de la señal
    % t = vector de tiempos
function [y_t] = mod_am_BLU_hilbert(x_t,Fw,fo,Fs,t)

    % Función para la obtención de un filtro tipo Hilbert especificando el orden
    H = designfilt('hilbertfir','FilterOrder',fo, ...
           'TransitionWidth',400,'SampleRate',Fs);

    % Aplicación del proceso de filtrado
    xp_t= filter(H,[x_t;zeros(fo/2,1)]);
```

```
    % Aplicación de esquema de modulación en doble rama
    xp_t= xp_t(fo/2+1:end);
    y1_t = x_t.*cos(2*pi*Fw*t);
    y2_t = xp_t.*sin(2*pi*Fw*t);

    y_t = y1_t+y2_t; % Señal modulada

end
```

En este caso, la representación matemática para módulo del espectro de la señal modulada es idéntica al caso anterior de modulación en banda lateral única, salvo por la respuesta en frecuencia del filtro empleado. Siendo $HL(f)$ la notación para expresar la transformada de Fourier de la señal $hl(t)$, se tiene:

$$HL(f) = j \quad \forall f < 0$$

$$HL(f) = -j \quad \forall f > 0$$

siendo j el término empleado para denotar a la unidad imaginaria.

La forma del espectro de la señal modulada con esta técnica es idéntica a la representada en la Ilustración 22, pero puede suponer una reducción en la capacidad de computación necesaria para el procesamiento de la señal, especialmente si el método equivalente emplea filtros digitales de un orden relativamente elevado.

5.2.3 MÉTODO DE FILTRADO

El procesado de la señal de voz empleada como vector de ataque requiere de la aplicación de métodos de filtrado en partes concretas del proceso de modulación.

Por una parte, puede ser necesario un procesamiento previo de la señal de voz, especialmente crítico si esta se ha obtenido mediante un sistema de grabación de una voz natural, para eliminar o, como mínimo, reducir el nivel de ruido situado por encima de las frecuencias

propias de la voz humana. Por otra parte, también se requiere del uso de filtros digitales en la implementación de la técnica de modulación en banda lateral única con un filtro básico.

Dentro del filtrado digital, encontramos principalmente dos tipos de filtros: FIR e IIR.

5.2.3.1 Filtros FIR

Los filtros FIR se caracterizan por tener una respuesta al impulso finita y carecer de rama de retroalimentación. Dentro de las ventajas de uso de este tipo de filtros, cabe destacar su estabilidad (siempre son estables) y la posibilidad de generar respuestas lineales en fase. Son especialmente demandados en aplicaciones donde la linealidad de la fase de la señal filtrada es crítica. Como característica negativa, es reseñable la relación entre su orden y su efectividad: conseguir un filtro FIR que se aproxime lo suficientemente bien a un filtro ideal puede requerir un orden de filtro demasiado elevado para las características de la aplicación en la que se vaya a implementar.

5.2.3.2 Filtros IIR

En cuanto a los filtros IIR, estos presentan una respuesta al impulso infinita y una rama de retroalimentación. Su principal ventaja frente a los filtros FIR es que permiten conseguir las mismas prestaciones de filtrado consumiendo menos tiempo y memoria, de manera que son especialmente demandados en aplicaciones con limitación de recursos o que requieren de una respuesta en tiempo real (Advsolned, 2020).

5.2.3.3 Selección de filtro

Para el caso de nuestro modelo de ataque, el tiempo consumido en el proceso de filtrado no es un elemento crítico, en cuanto a que la emisión de la señal de ataque se corresponde con una fase independiente a su procesado y, por tanto, temporalmente desvinculada. Además, al desconocer a priori la influencia que alteraciones no lineales sobre la fase de la señal de voz puede tener sobre la efectividad del ataque, es especialmente interesante trabajar con un tipo de filtro que nos asegure una respuesta lineal en fase.

Por ello, y teniendo también en cuenta como punto positivo la seguridad que nos ofrecen este tipo de filtros en cuanto a su estabilidad, se ha decidido aplicar filtros digitales de tipo FIR en los procesos que así lo requieran dentro del modelo de ataque.

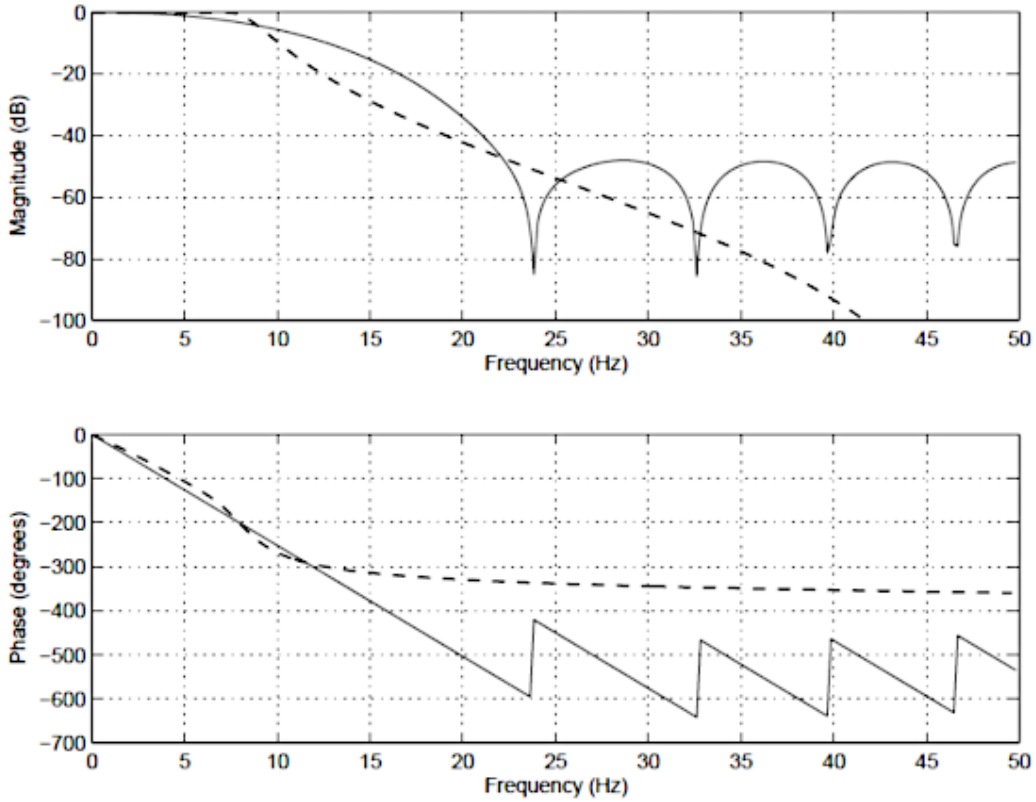


Ilustración 23. Respuesta en frecuencia de un filtro FIR de orden 14 (línea sólida) y un filtro IIR de orden 4 (línea punteada). Fuente: (Advsolned, 2020)

La respuesta en frecuencia de este tipo de filtros no deja de ser una aproximación a una respuesta en frecuencia deseada para un filtro ideal, tal y como se puede observar en la Ilustración 23. En el marco de esta aproximación, hay tres elementos críticos que pueden servir para representar el grado de exactitud de la misma: rizado en la banda de paso (idealmente nulo), ancho de la banda de transición (idealmente nulo) y atenuación en la banda de corte (idealmente infinita). Se ha optado por tener control directo sobre estas tres magnitudes, sacrificando el control sobre el orden del filtro, para así poder adaptar las

características del mismo a las necesidades del modelo de ataque, de acuerdo con los resultados experimentales. El objetivo es poder llegar a un compromiso entre el orden del filtro (capacidad de computación requerida) y la efectividad del ataque.

5.2.4 SISTEMA EMISOR DE AUDIO

El sistema emisor será el encargado de reproducir la señal de ataque, una vez esta haya sido sometida a todos los procesos previos necesarios, y "lanzarla" contra el dispositivo objetivo, buscando activar el mecanismo del software de reconocimiento de voz.

En concordancia con los objetivos expuestos, el sistema emisor debe de estar compuesto por dispositivos que puedan ser fácilmente encontrados en un entorno comercial o que sirvan como representación generalizada del mercado de dispositivos emisores de audio de uso común.

Se ha optado por usar sistemas de altavoces para emisión de audio en estéreo. Se emplearán dos pares de altavoces de distinta gama y ganancia para poder establecer una comparación entre las prestaciones de los mismos. Sin embargo, ambos sistemas de audio presentan una especificación que supone un estándar de facto en el mercado de los altavoces de uso genérico, asegurando una respuesta en frecuencia lineal hasta un valor máximo de frecuencia de 20 kHz.

La posible manifestación de la característica no lineal de estos sistemas en frecuencias superiores a la señalada, así como el efecto directo que este hecho podría tener sobre la viabilidad del ataque será objeto de estudio durante la fase de pruebas.

5.2.5 DISPOSITIVO OBJETIVO DEL ATAQUE

El dispositivo objetivo será el elemento sobre el cual se probará directamente el modelo de ataque diseñado para comprobar así su efectividad. Como requisito principal, se debe tratar de algún tipo de dispositivo electrónico en cuyo sistema operativo se encuentre configurada una herramienta software de reconocimiento de voz.

Para maximizar el alcance de las generalidades que pueden ser extraídas de los resultados del estudio, se ha decidido emplear diferentes dispositivos, tratando de abarcar:

- Dispositivos de diferentes generaciones
- Algoritmos de reconocimiento de voz de distintos desarrolladores
- Sistemas de reconocimiento de voz con y sin requerimiento de señal de activación.
- Dispositivos de diferente morfología, con especial interés en la existencia de variedad en cuanto al sistema de entrada audio.

El modelo de ataque presupone que el atacante no tiene acceso al dispositivo y por tanto este no puede ser objeto de ningún tipo de manipulación antes o durante el ataque.

Capítulo 6. ANÁLISIS DE RESULTADOS

6.1 DISEÑO DE PRUEBAS

Los objetivos de las pruebas que a continuación se detallan serán:

- Estudiar las características fundamentales de los dispositivos objetivo empleados, buscando similitudes y diferencias entre ellos, así como datos que permitan conocer mejor su esquema de funcionamiento.
- Probar la efectividad del modelo de ataque diseñado, basado en la hipótesis de la característica no lineal de los componentes electrónicos básicos del dispositivo receptor.

6.1.1 PRUEBA DE UMBRAL DE POTENCIA

Descripción del objetivo de la prueba

El objetivo de esta prueba es establecer el umbral de potencia mínimo necesario para activar el mecanismo de reconocimiento de voz del dispositivo.

Tal como referencia el punto 2.1, el procesamiento de una señal acústica por parte de un software de reconocimiento de discurso viene precedido por una fase en la que se aplican una serie de filtros rudimentarios al fragmento de audio. Estos filtros siguen criterios de corte basados en valores de potencia mínima y frecuencia máxima. Por tanto, conocer la potencia media mínima que debe portar una señal para superar este filtro inicial y pasar a la fase de modelado matemático puede ser crítico de cara a posibles modificaciones en el modelo de ataque.

A priori, se desconoce si la característica no lineal del receptor puede traducirse de manera directa en cambios en la amplitud de la señal recibida. No obstante, sí es conocido el efecto de atenuación intrínseco a la propagación de las ondas acústicas por el medio aéreo. Por ello,

se espera que la señal recibida por el dispositivo objetivo se caracterice por un valor de potencia menor que la señal original que es empleada como vector de ataque.

Dado que las diferentes técnicas de modulación en amplitud a emplear implican una reducción de la energía de la señal original, el umbral de potencia mínimo que se debe superar para que la señal emitida sea reconocida por el dispositivo objetivo puede ser un factor limitante para el uso de algún tipo de modulación o para la distancia máxima a la cual el ataque es efectivo.

Por último, es interesante estudiar si estos dispositivos establecen un umbral de potencia de valor único o si este umbral es adaptativo y dependiente del tipo de señal esperada (activación o comando de actuación).

Descripción del desarrollo de la prueba

Para el desarrollo de esta prueba, se dispondrá de señales de actuación (y activación, si así se requiere) obtenidas a partir de una grabación de voz humana para cada uno de los dispositivos objetivo. Una vez preparadas las señales, se seguirán los pasos siguientes:

1. Las señales serán procesadas en banda base y la única modificación que experimentarán será un proceso de filtrado básico con un filtro paso-bajo de frecuencia de corte igual a 8 kHz. El motivo del filtrado es eliminar así efecto del ruido de alta frecuencia que pueda estar presente en la muestra de audio original por motivo de las condiciones de grabación.
2. Utilizando el conjunto de altavoces de gama más alta, se emitirá la señal, a máxima potencia y a una distancia controlada del dispositivo objetivo.
3. Periódicamente, se irá reduciendo el nivel de potencia de la señal, hasta registrar el valor para el cual el dispositivo deja de responder correctamente al comando enunciado en la señal de audio. Se considerará que el dispositivo deja de responder correctamente a la señal cuando no ejecute la acción asociada al comando empleado durante tres intentos consecutivos.

4. Se repetirá este mismo proceso para diferentes distancias de emisión y diferentes dispositivos. Para cada dispositivo, se registrará el valor mínimo de potencia para el cual es capaz de interpretar la señal en cada una de las distancias probadas.

6.1.2 PRUEBA DE DESVIACIÓN EN FRECUENCIA

Descripción del objetivo de la prueba

El objetivo de esta prueba es determinar el grado máximo de desviación que puede presentar una señal en banda base con respecto a la frecuencia central $f_0 = 0\ Hz$ para seguir siendo reconocida por el dispositivo asistente de voz.

El hecho de no disponer de un modelo matemático para representar la característica no lineal de los componentes del receptor nos impide, a priori, conocer qué nuevas frecuencias serán generadas cuando la señal atraviese el sistema de captación de audio (el posible efecto de una respuesta no lineal se aborda en el apartado Capítulo 5.). En algunos casos, es posible que, por efecto de una respuesta en frecuencia no lineal, se genere una réplica de la señal original a partir de la señal modulada (en este caso, estaría teniendo lugar un proceso de "demodulación indirecta" de la señal, al que se alude en el apartado 3.5)

Sin embargo, puede darse la situación de que la réplica de la señal que se genera en banda base no se encuentre exactamente centrada en $f_0 = 0\ Hz$, sino que experimente una cierta desviación. En ese caso, la señal generada sería la equivalente a modular la señal en banda base con una frecuencia de portadora de unos pocos hercios.

Resulta de interés estudiar si estas ligeras desviaciones en frecuencia harían que una señal dejase de ser comprendida por un determinado asistente de voz. Cuanto mayor sea la tolerancia intrínseca del sistema a esta desviación en frecuencia, mayores son las probabilidades de que el ataque resulte efectivo para un cierto rango de frecuencias.

Descripción del desarrollo de la prueba

Para el desarrollo de esta prueba, se dispondrá de señales de actuación (y activación, si así se requiere) obtenidas a partir de una grabación de voz humana para cada uno de los dispositivos objetivo. Una vez preparadas las señales, se seguirán los pasos siguientes:

1. Las señales serán previamente filtradas y posteriormente moduladas en amplitud. Se empleará un tipo de modulación en banda lateral única superior (explicada en apartado 5.2.2.3), debido a que el resto de métodos de modulación ocasionarían solapamiento espectral entre las réplicas de la señal si se usa una frecuencia de portadora de unos pocos cientos de hercios.
2. Partiendo de la señal centrada en $f_0 = 0\,Hz$, se irá aumentando la frecuencia de portadora en escalones de 50 Hz
3. Se registrará el valor de frecuencia de portadora (desviación máxima) para el cual dispositivo deja de responder correctamente al comando enunciado en la señal de audio, aplicando la misma consideración que en la prueba anteriormente descrita.
4. Se repetirá este mismo proceso para diferentes distancias de emisión y diferentes dispositivos.

6.1.3 PRUEBA DE EFECTIVIDAD DEL ATAQUE

Descripción del objetivo de la prueba

El objetivo de esta prueba es comprobar la efectividad del modelo de ataque diseñado. Una muestra de la efectividad del ataque serviría a su vez para confirmar la validez de la hipótesis central de trabajo, expuesta en el punto Capítulo 5. .

Para esta prueba, es necesario contemplar el factor de no linealidad en los dispositivos emisores de audio. Si este efecto no lineal se asemeja al descrito en el apartado Capítulo 5. , puede llegar a ser imposible emitir una señal no audible mediante el uso de estos aparatos.

Por tanto, primero se debe comprobar si es posible emitir una señal acústicamente inapreciable para un ser humano usando alguno de los dispositivos emisores de los que se dispone para la prueba.

Descripción del desarrollo de la prueba

Para el desarrollo de esta prueba, se tomará una señal cualquiera de audio dentro del banco de señales con las que se ha trabajado en pruebas anteriores. Seguidamente:

1. Empleando una frecuencia de portadora de valor inicial 20 kHz e incrementando este valor progresivamente hasta los 30 kHz, se someterá a las diferentes técnicas de modulación expuestas en el apartado 5.2.2.
2. El resultado de la prueba será positivo si, para alguna de las frecuencias de portadora empleadas, la señal emitida por los altavoces resulta inaudible para un oyente imparcial. En caso contrario, el resultado se considerará negativo.

En caso de que el resultado de esta prueba sea positivo y se compruebe que es posible emitir una señal no audible con el equipo del que se dispone, la siguiente prueba a realizar se encontrará especialmente dedicada a ponderar la efectividad del ataque:

1. Como en las pruebas anteriores, se dispondrá de señales de actuación (y activación, si así se requiere) obtenidas a partir de una grabación de voz para cada dispositivo.
2. Para los rangos de frecuencias en los que se trabaje con una señal inaudible, estas señales serán procesadas con los métodos de modulación que se hayan demostrado efectivos.
3. Cada una de estas señales procesadas se emitirá 10 veces y se registrará el resultado en forma de porcentaje de ataques efectivos, tomando como ataque efectivo aquel en el que el dispositivo objetivo ejecute el comando asociado a la señal original.

En caso de obtener un resultado negativo en la prueba de emisión de señales inaudibles, las pruebas posteriores irán encaminadas a obtener más información sobre la característica no lineal de los dispositivos emisores:

1. Inicialmente, se seleccionarían aquellas franjas de frecuencia y métodos de modulación para los cuales sea posible distinguir la señal original en banda base entre los diferentes componentes de frecuencia de la señal emitida.

2. Seguidamente, se reproduciría la prueba de efectividad del ataque de forma análoga al caso anteriormente expuesto, comprobándose ahora el porcentaje de éxito con el que el dispositivo objetivo reconoce la réplica en banda base de la señal original generada por la característica no lineal del emisor.
3. La imposibilidad de reproducir una señal inaudible implicaría que el efecto no lineal de los componentes electrónicos del dispositivo emisor estaría generando réplicas de la señal en frecuencias audibles.

6.2 *DESARROLLO DE LAS PRUEBAS Y OBTENCIÓN DE RESULTADOS*

Tras ejecutar la batería de pruebas planteada, de acuerdo con la descripción ofrecida en el apartado Capítulo 6. , se han obtenido una serie de resultados que permiten la elaboración de un conjunto de conclusiones.

Todas las pruebas se han desarrollado en una habitación silenciosa, con ruido ambiente no superior a los –50 dBW. La Ilustración 24, muestra una medida del ruido ambiente del entorno de pruebas tomada con una aplicación móvil emuladora de un espectrómetro.

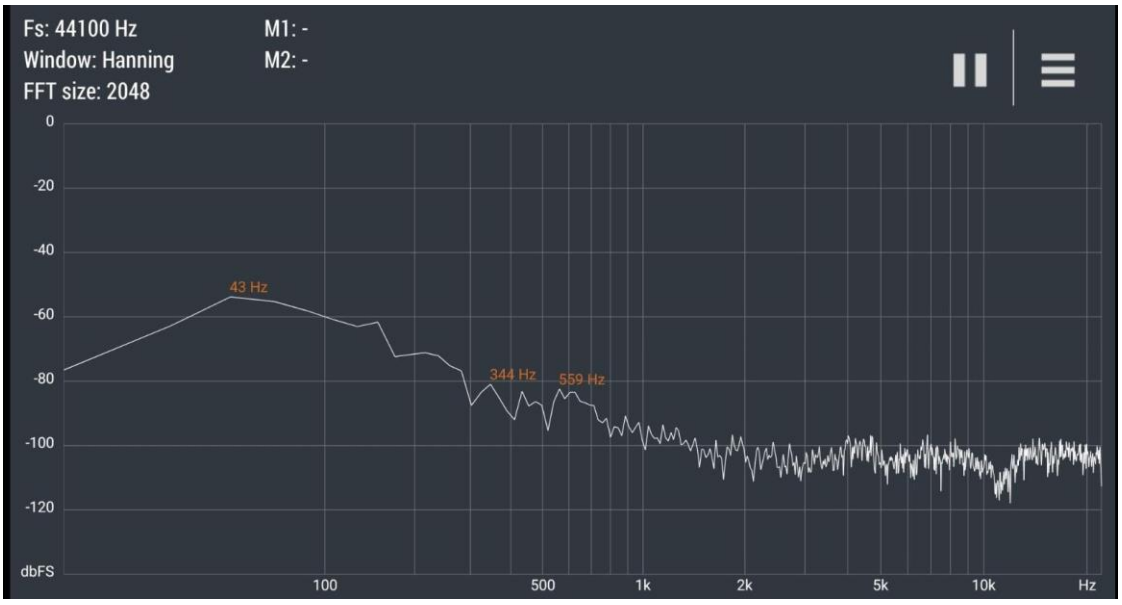


Ilustración 24. Medida de intensidad del ruido ambiente en el entorno de pruebas

6.2.1 PRUEBA DE UMBRAL DE POTENCIA

6.2.1.1 Dispositivos empleados y especificaciones de la prueba

Dispositivo	*Asistente de voz*	*Desarrollador*	*Año de lanzamiento*
Amazon Echo (2ª generación)	Alexa	Amazon	2016
Ipad Mini 2	Siri	Apple	2014

Tabla 2. Dispositivos receptores empleados en la prueba del umbral de potencia

La Tabla 2 muestra los dos dispositivos asistentes de voz seleccionados para la prueba. Ambos requieren de una señal de activación para acceder el resto de comandos de voz disponibles. Esto hace posible estudiar las diferencias entre el umbral mínimo de potencia necesario que debe superar una señal para activar el dispositivo y para ejecutar algún comando de actuación.

Además, el hecho de haber sido manufacturados por desarrolladores distintos posibilita una comparación directa entre las prestaciones de ambos.

Especificaciones

Se han tomado medidas del umbral mínimo de potencia necesario para provocar una respuesta de cada uno de los sistemas asistentes de voz escogidos en dos escenarios diferentes: La especificación "*Partiendo de inactividad"* hace referencia al uso del comando de activación propio de cada sistema, mientras que "*Partiendo de escucha activa"* implica el uso de un comando de actuación sobre el dispositivo previamente activado y en estado de espera.

El experimento se ha repetido para 5 distancias diferentes, medidas entre el dispositivo objetivo y el sistema de altavoces encargado de emitir la señal de audio.

Los valores obtenidos hacen referencia a la potencia de la señal emitida multiplicada por la ganancia máxima de los altavoces empleados.

En este caso, la ganancia de los altavoces tiene un valor de $G = 2.5$, mientras que la potencia de la señal emitida ha sido calculada aplicando el sumatorio de los cuadrados absolutos de las muestras de la señal en el dominio de tiempo, divididas entre la longitud de la señal. De forma equivalente, se puede considerar la aplicación del cálculo del cuadrado de su nivel de RMS., mostrada en la Ecuación 12

Ecuación 12
$$P(Watts) = 2.5 \frac{\sum_{n=1}^{n=L} x[n]^2}{L}$$

6.2.1.2 Resultados

Distancia (cm)	*Partiendo de escucha activa*		*Partiendo de inactividad*	
	W	dBW	W	dbW
20	1,4513E-07	-68,38	1,7535E-07	-67,56
40	8,3575E-09	-80,78	8,6550E-09	-80,63
60	9,8175E-09	-80,08	1,0818E-08	-79,66
80	3,0875E-08	-75,10	3,8775E-08	-74,11
100	8,7800E-08	-70,57	1,1885E-07	-69,25

Tabla 3. Resultados de la prueba de umbral de potencia para el dispositivo Amazon Echo

Distancia (cm)	*Partiendo de escucha activa*		*Partiendo de inactividad*	
	W	dBW	W	dbW
20	1,8805E-08	-77,3	1,4375E-07	-69,03
40	2,4143E-08	-76,2	3,6250E-07	-64,71
60	2,5450E-08	-75,9	6,3500E-07	-62,15
80	3,5450E-08	-74,5	7,9450E-07	-61,20
100	7,9875E-08	-71,0	8,3175E-07	-61,24

Tabla 4. Resultados de la prueba de umbral de potencia para el dispositivo Ipad Mini 2

6.2.1.3 Conclusiones

Relación entre la potencia umbral y la distancia al emisor

Tras representar los valores de potencia mínimos registrados frente a la distancia a la cual fueron tomados (Ilustración 25), es posible determinar que existe una relación entre la distancia y la potencia mínima de emisión necesaria para activar el dispositivo en cada caso.

De forma genérica se puede enunciar que, **a mayor distancia del dispositivo objetivo, mayor es la potencia con la que es necesario emitir la señal** para superar la fase de filtrado previo realizada por el subsistema de captación de audio del dispositivo objetivo. Se trata de un resultado esperado, pues pone de manifiesto el efecto atenuante del medio sobre la propagación de las ondas de sonido.

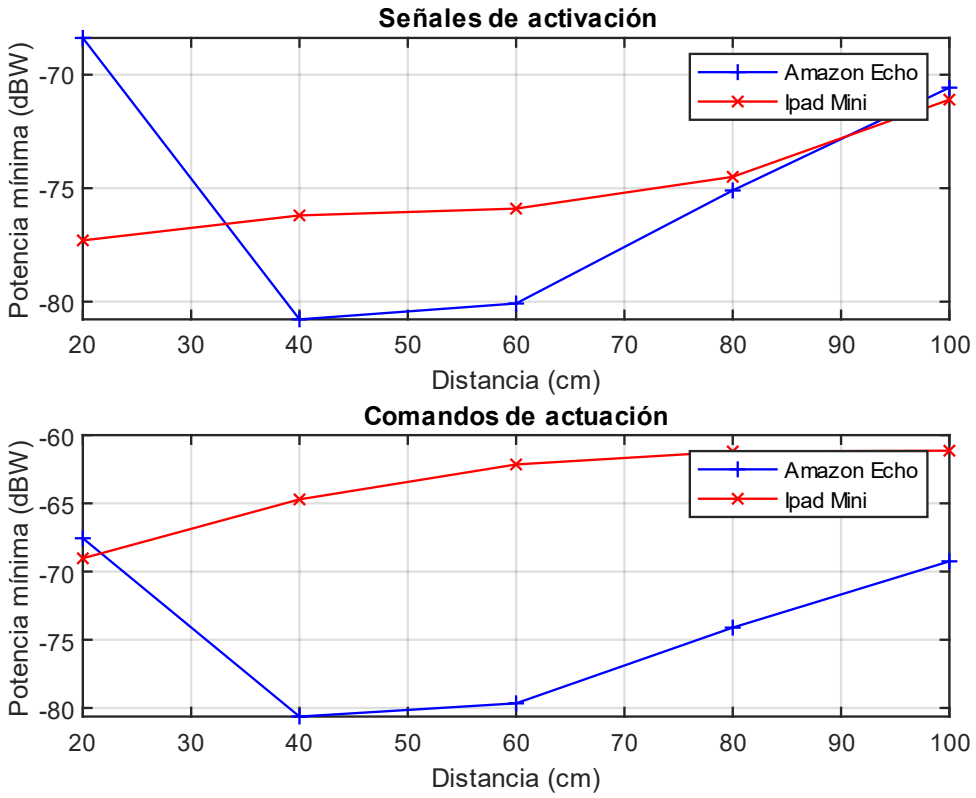


Ilustración 25. Representación gráfica de los resultados obtenidos en la prueba de umbral de potencia. Relación entre potencia y distancia.

Por otra parte, la comparación de los datos obtenidos para dispositivos distintos muestra algunas diferencias significativas entre los productos de diferentes desarrolladores:

- En primer lugar, el dispositivo Amazon Echo presenta un dato atípico para la distancia de 20 cm, el cual no se adecúa a la progresión lineal que parecen seguir los datos tomados para el resto de distancias. Este hecho ocurre tanto para el comando de activación como para el de actuación y no se manifiesta en los resultados obtenidos para el dispositivo Ipad Mini.
- En segundo lugar, aunque los valores de potencia mínima parecen seguir una progresión ascendente en ambos casos, la ratio de crecimiento de esta gráfica se adivina notablemente mayor para el dispositivo Amazon Echo. Por algún motivo, el

nivel de potencia de señal registrado por este dispositivo presenta una mayor sensibilidad a la distancia de emisión de la señal.

Estos resultados sugieren que **existe una influencia directa de la morfología del dispositivo, así como de la localización de los micrófonos** destinados a la captación de señales de audio para el mismo, sobre la potencia mínima de emisión necesaria para ejecutar un comando de voz.

Influencia de la morfología del dispositivo

El dispositivo Amazon Echo presenta una geometría cilíndrica que sugiere una distribución de los elementos de captación de audio en los 360º que abarcan la longitud de la circunferencia de su base. Esta distribución explicaría el dato atípico que aparece en el estudio del umbral de frecuencia para el menor valor de distancia.

Debido a la geometría del dispositivo objetivo y la forma de propagación de la señal de audio como una onda tridimensional, la distancia entre emisor y receptor afecta al área del dispositivo receptor expuesta directamente a las ondas acústicas del mensaje, siendo esta área de contacto directamente proporcional a la distancia entre ambos. Para un valor de distancia suficientemente pequeño, puede ser que este efecto tenga mayor peso que la atenuación de la onda en el medio. Consecuentemente, es posible explicar que para un receptor con este tipo de morfología se requiera mayor potencia de emisión de la señal a 20 cm que a 40 cm.

Por otra parte, el dispositivo Ipad Mini presenta su sistema de captación de audio en un único punto de entrada de sonido, lo cual explica que todos los datos recogidos se adecúen a la progresión lineal esperada, pues la geometría del receptor no juega ningún papel reseñable en este caso.

La posible influencia de la morfología del dispositivo en la prueba queda representada en la Ilustración 26.

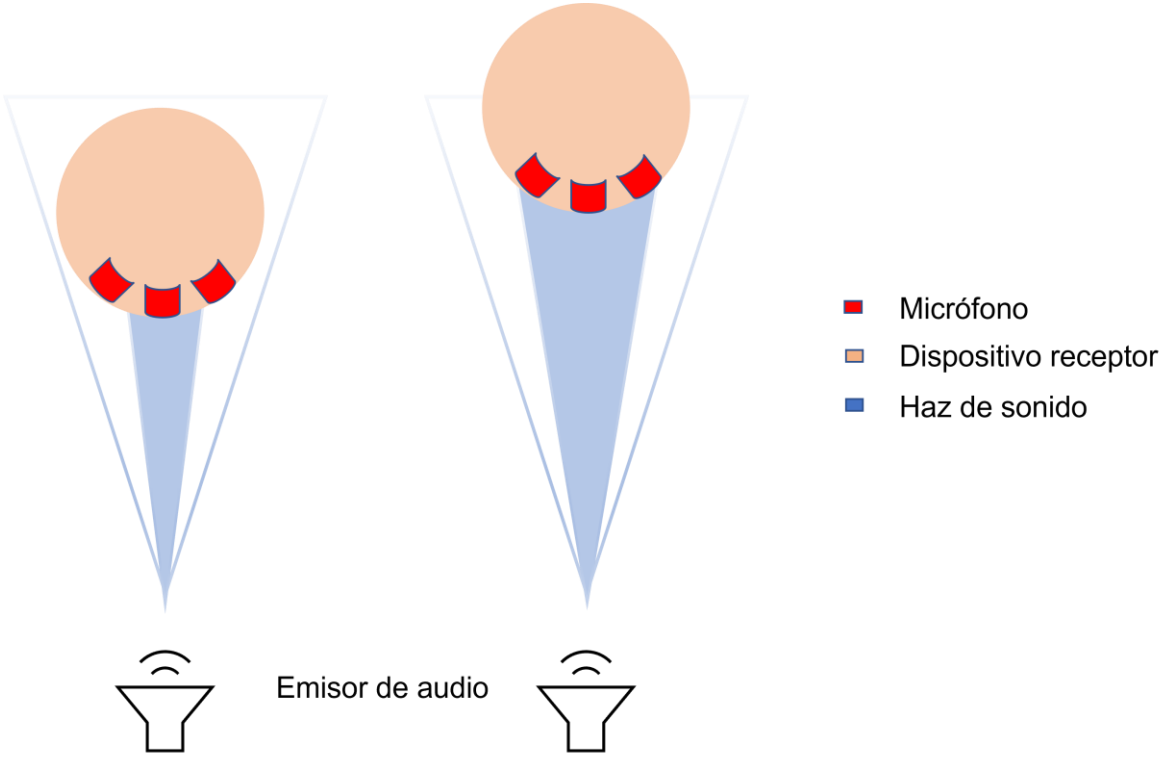


Ilustración 26. Influencia de la morfología del dispositivo en el resultado de la prueba de umbral de potencia

De acuerdo con esta línea de razonamiento, es posible explicar también la diferencia en la sensibilidad de la potencia registrada por cada dispositivo a la distancia entre emisor y receptor en base al área activa de captación de audio. Entendiendo por área activa de captación de audio aquella porción de superficie del dispositivo que cuente con algún tipo de micrófono destinado a la captación de señales acústicas, es posible enunciar que, **a mayor superficie activa de captación de audio, más sensible será el dispositivo** al efecto que la distancia de propagación de la señal tiene sobre el proceso de reconocimiento de la misma.

Relación entre la potencia umbral y el tipo de comando

En cuanto a la comparación entre tipos de comandos distintos para un mismo dispositivo, en ambos casos se aprecia una diferencia entre el nivel mínimo de potencia requerido por un comando de activación y uno de actuación.

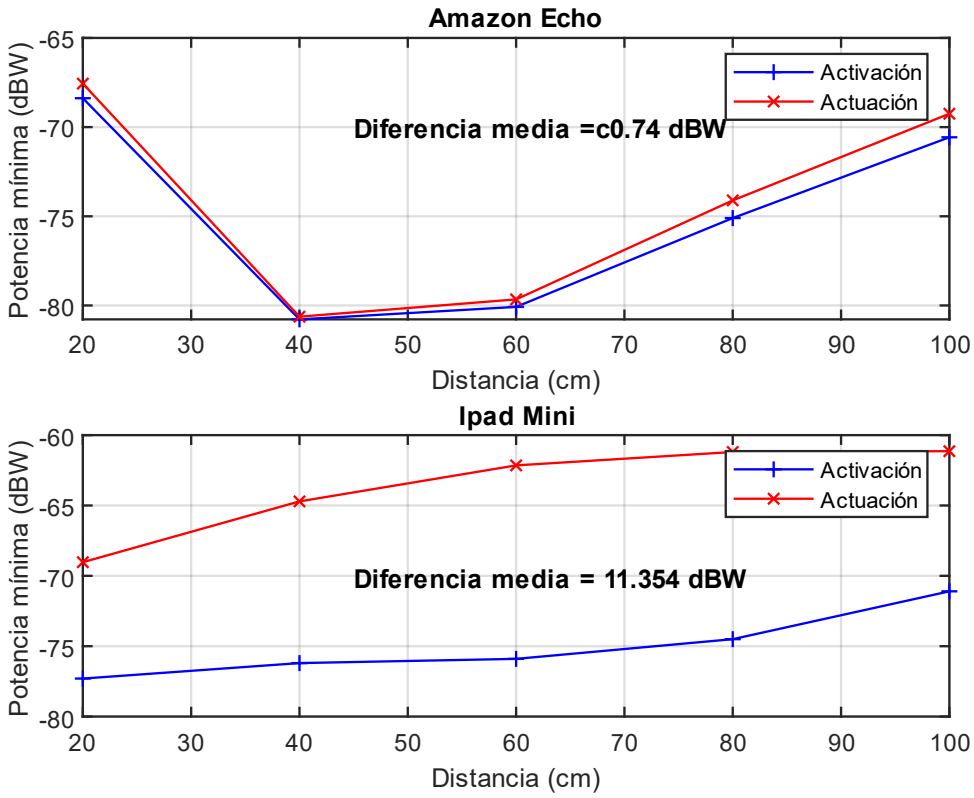


Ilustración 27. Representación gráfica de los resultados obtenidos en la prueba de umbral de potencia. Relación entre potencia y tipo de comando.

Esta observación permite deducir un fundamento importante del funcionamiento de los dispositivos asistentes de voz: **las señales de activación requieren ser recibidas con un nivel de potencia mínimo más elevado que las señales de actuación**. Probablemente, la lógica detrás de este criterio de diseño se sustente en minimizar la posibilidad de que estos dispositivos se activen de manera errónea.

Una vez que el dispositivo está activado, lo más probable es que el usuario se disponga a hacer uso de algún comando de voz, por lo que tiene sentido que en este caso el umbral de potencia disminuya, a la espera de recibir dicho comando.

La diferencia media entre umbrales de potencia es diferente para cada dispositivo, siendo mayor en el caso del Ipad Mini. En este caso, se trata de un rasgo establecido por la configuración de los filtros de potencia internos a cada dispositivo, por lo que se atribuye a criterio del desarrollador.

6.2.2 PRUEBA DE DESVIACIÓN EN FRECUENCIA

6.2.2.1 Dispositivos empleados y especificaciones de la prueba

Dispositivo	*Asistente de voz*	*Desarrollador*	*Año de lanzamiento*
Amazon Echo (2ª generación)	Alexa	Amazon	2016
Ipad Mini 2	Siri	Apple	2014
Samsung Galaxy S21 *	Bixby	Samsung	2021
Samsung Galaxy S21 *	Google Assistant	Google	2021

Tabla 5. Dispositivos receptores empleados en la prueba de desviación de potencia

*el dispositivo ha pasado las pruebas variando la entrada de audio

Para esta prueba, se ha decidido emplear 4 asistentes de voz de diferentes desarrolladores, con el objetivo de establecer una comparación entre ellos (mostrados en la Tabla 5). Además, se ha considerado interesante incluir un nuevo dispositivo con una fecha de desarrollo y lanzamiento más reciente, para estudiar si existe algún tipo de distinción en la respuesta de este ante las señales de prueba frente a la respuesta ofrecida por dispositivos más antiguos.

Como última consideración, para el dispositivo Samsung Galaxy S21, se ha determinado que las pruebas se desarrollen por duplicado. En cada caso, se empleará una entrada de audio distinta, siendo las dos alternativas la entrada de audio propia del dispositivo y la entrada de audio de unos auriculares externos conectados al dispositivo. La variación de la entrada de audio nos permitirá determinar el grado de influencia del subsistema de captación de audio en los resultados de esta prueba.

Especificaciones

En este caso, las señales de prueba se corresponden con señales moduladas en amplitud empleando el método de banda lateral única superior. La señal original empleada es la misma para todos los dispositivos y se corresponde con un comando que genera una respuesta común en todos ellos.

Los resultados de la prueban registran la frecuencia máxima de la señal portadora empleada en la modulación a partir de la cual el comando de voz deja de ser entendible por el dispositivo. Para determinar que un comando ha dejado de ser entendible, el criterio empleado ha sido determinar cómo no entendible cualquier comando que recibe una respuesta negativa (el dispositivo objetivo no ejecuta la acción asociada al comando o ejecuta una acción diferente a la esperada) en tres intentos consecutivos.

Para cada dispositivo, la prueba ha sido repetida en 5 ocasiones, variando la distancia entre el receptor y el sistema emisor de audio en cada una de ellas. Para cada distancia, se registra la frecuencia máxima de portadora (en hercios) y la potencia de la señal empleada en cada caso (en dbW).

6.2.2.2 Resultados

Dispositivo: Amazon Echo 2ª generación

Distancia (cm)	Frecuencia de portadora (Hz)	Potencia de la señal (dbW)
20	350	-31,30
40	350	-31,30
60	350	-31,30
80	320	-31,31
100	320	-31,31

Tabla 6. Resultados de la prueba de desviación en frecuencia para el dispositivo Amazon Echo

Dispositivo: Ipad Mini 2

Distancia (cm)	Frecuencia de portadora (Hz)	Potencia de la señal (dbW)
20	340	-27,70
40	320	-27,62
60	320	-27,50
80	300	-27,38
100	300	-27,47

Tabla 7. Resultados de la prueba de desviación en frecuencia para el dispositivo Ipad Mini 2

Dispositivo: Samsung Galaxy S21s / Bixby

Distancia (cm)	Frecuencia de portadora (Hz)	Potencia de la señal (dbW)
20	300	-26,52
40	300	-26,35
60	310	-26,52
80	300	-26,35
100	280	-26,11

Tabla 8. Resultados de la prueba de desviación en frecuencia para el dispositivo Samsung Galaxy S21s / Bixby

Dispositivo: Samsung Galaxy S21s / Bixby (con auriculares)

Distancia (cm)	Frecuencia de portadora (Hz)	Potencia de la señal (dbW)
20	80	-24,31
40	80	-24,88
60	90	-24,31
80	90	-24,31
100	100	-24,28

Tabla 9. Resultados de la prueba de desviación en frecuencia para el dispositivo Samsung Galaxy S21s / Bixby usando auriculares

Dispositivo: Samsung Galaxy S21s / Google Assistant

Distancia (cm)	Frecuencia de portadora (Hz)	Potencia de la señal (dbW)
20	300	-26,35
40	300	-26,35
60	300	-26,35
80	300	-26,35
100	300	-26,36

Tabla 10. Resultados de la prueba de desviación en frecuencia para el dispositivo Samsung Galaxy S21s / Google Assistant

Dispositivo: Samsung Galaxy S21s / Google Assistant (con auriculares)

Distancia (cm)	Frecuencia de portadora (Hz)	Potencia de la señal (dbW)
20	200	-24,61
40	220	-24,98
60	240	-25,88
80	240	-25,88
100	240	-25,88

Tabla 11. Resultados de la prueba de desviación en frecuencia para el dispositivo Samsung Galaxy S21s / Google Assistant usando auriculares

6.2.2.3 Conclusiones

Relación entre la desviación tolerada y el tipo de dispositivo

La comparación entre asistentes de voz pone de manifiesto cierta variabilidad en los valores de los resultados obtenidos. La Tabla 12 muestra la media aritmética y desviación típica de los resultados obtenidos para cada uno de los dispositivos.

Media			
Alexa	**Siri**	**Bixby**	**Google**
332 Hz	316 Hz	298 Hz	300 Hz
Desviación típica			
Alexa	**Siri**	**Bixby**	**Google**
16.4 Hz	16.7 Hz	10.9 Hz	0 Hz

Tabla 12. Media aritmética y desviación típica de los resultados de la prueba de desviación en frecuencia para cada dispositivo asistente

Se aprecia como el asistente que tolera una mayor desviación en frecuencia de la señal emitida en banda base es *Alexa*, 34 Hz superior al menor valor medio registrado atribuido a *Bixby*.

También resulta destacable que el asistente de Google parece manifestar una tolerancia muy alta a la variación de distancia, pues la desviación en frecuencia aceptada por este asistente de voz es idéntica para las 5 distancias de medida registradas.

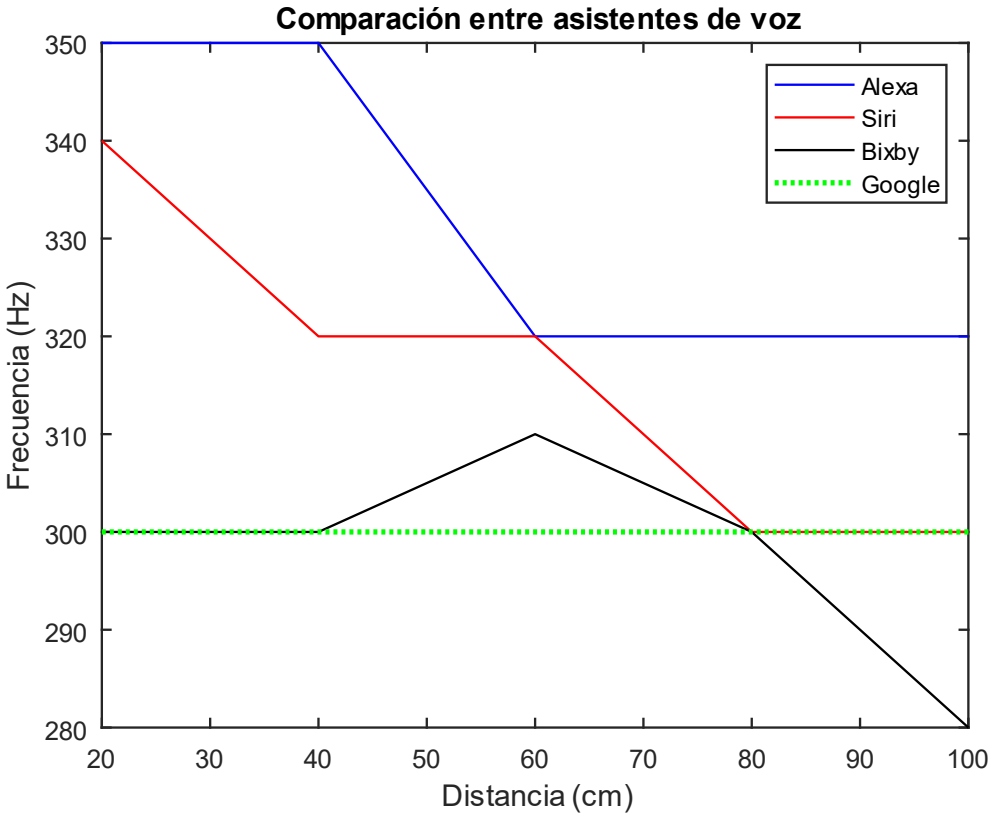


Ilustración 28. Representación gráfica de los resultados obtenidos en la prueba de desviación en frecuencia. Relación entre frecuencia y distancia.

Aunque la diferencia entre la respuesta de los diferentes asistentes no es notable, sí que cabe plantearse si el motivo de la variabilidad registrada se relaciona con aspectos *hardware*, característicos del dispositivo en sí, o *software*, los cuales tendrían relación directa con el algoritmo de interpretación de audio usado por el asistente.

Los resultados representados en la Ilustración 28 invitan a pensar que se trata de una combinación de ambos factores. Este hecho resulta especialmente notable al establecer la comparación entre los asistentes instalados en el mismo dispositivo (*Bixby* y *Google Assistant):* por un lado, se observa que presentan un valor medio de desviación en frecuencia tolerada muy similar entre ellos y diferente si lo comparamos con los otros dos dispositivos; por otro lado, uno de ellos muestra una mayor tolerancia a la variación de distancia con

respecto al emisor de la señal. Mientras que el dato del valor medio pone de manifiesto una diferencia entre dispositivos, el dato de la desviación típica se relaciona con una diferencia entre asistentes.

La diferencia entre dispositivos deriva del uso de diferentes elementos *hardware* (micrófonos, amplificadores, filtros…) usados para manufacturar cada uno de ellos, así como de la morfología particular que cada uno presenta.

La diferencia entre asistentes se debe, sin embargo, a una cuestión puramente programática, derivada del tipo de modelo que el asistente use para caracterizar las muestras de voz y del algoritmo que esté siendo empleado para comparar dichas muestras con modelos de comandos conocidos.

Efecto de la distancia

En cuanto al efecto de la distancia sobre el resultado de la prueba**, se puede comprobar la existencia de una relación inversa entre el valor de desviación media en frecuencia tolerado por los dispositivos y la distancia** a la que sitúa el par de altavoces en cada caso, tal como se muestra en la siguiente la Tabla 13:

Distancia	*20 cm*	*40 cm*	*60 cm*	*80 cm*	*100 cm*
Desviación media tolerada por los dispositivos	322.5 Hz	317.5 Hz	312.5 Hz	305 Hz	300 Hz

Tabla 13. Valor medio de la desviación en frecuencia tolerada por cada dispositivo para cada una de las distancias probadas

Efecto del uso de auriculares

El último objeto de estudio de esta prueba ha sido el impacto que tiene un cambio en la entrada de audio sobre el resultado de la misma. En este caso, lo resultados recogidos en la Tabla 14 sí muestran una diferencia notable:

	Bixby	*Google Assitant*
Sin auriculares	298 Hz	300 Hz
Con auriculares	88 Hz	228 Hz

Tabla 14. Comparación de los resultados obtenidos en la prueba de desviación en frecuencia según se usen o no auriculares

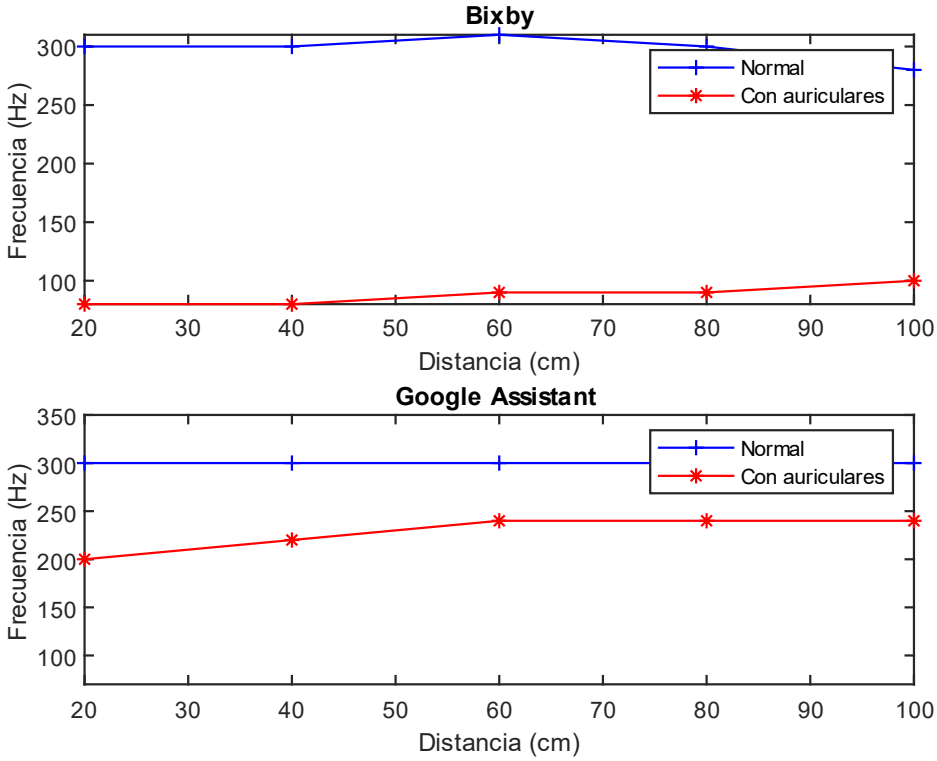


Ilustración 29. Representación gráfica de los resultados obtenidos en la prueba de desviación en frecuencia según se usen o no auriculares

El uso de auriculares reduce significativamente la desviación en frecuencia tolerada por el dispositivo, tal como se observa en la Ilustración 29. En este caso, este resultado es únicamente atribuible a la diferencia entre los componentes que constituyen ambos sistemas de captación de audio.

Las implicaciones de esta observación se pueden plantear desde dos enfoques diferentes:

- Desde un punto de vista de **versatilidad**, el hecho de que un sistema asistente de voz tolere una menor desviación en frecuencia de la señal en banda base con respecto a la frecuencia central nula se puede entender como un rasgo negativo. En este caso, el dispositivo sería susceptible de ser desbloqueado por un menor rango de frecuencias vocales.
- Desde un punto de vista de **seguridad**, este hecho puede suponer una ventaja en cuanto a que se reduce la posibilidad de efectividad de modelos de ataque como el estudiado en este proyecto. Un mayor rango admisible de desviación en frecuencia provocaría que el dispositivo estuviese más expuesto a que el posible efecto de la respuesta no lineal de sus componentes electrónicos generase una réplica en banda base que pudiese ser entendida por el algoritmo de reconocimiento de voz. En este sentido, el uso de una entrada de audio que disminuyese el valor de desviación en frecuencia máxima tolerada estaría protegiendo el dispositivo frente a este tipo de ataques.

De acuerdo con este último enfoque y teniendo en cuenta los datos obtenidos cuando cada dispositivo hace uso de su sistema de captación de audio estándar, el más seguro sería el *Samsung Galaxy S21*. Al tratarse del dispositivo con fecha de lanzamiento más reciente, se plantea la cuestión de si la industria tiene en cuenta el factor de seguridad asociado a la deriva en frecuencia tolerada como criterio para elección de los componentes electrónicos de los diseños más modernos o de si se trata de un resultado fortuito del proceso de fabricación.

6.2.3 PRUEBA DE REPRODUCCIÓN DE SEÑALES INAUDIBLES

6.2.3.1 Dispositivos empleados y especificaciones de la prueba

Modelo	*Desarrollador*	*Potencia RMS (W)*	*Rango de respuesta lineal (Hz)*
SBS 260	Creative	2.5	90 - 20000
Z120	Logitech	1.2	20 - 20000

Tabla 15. Dispositivos empleados en la prueba de audibilidad

Los dos sistemas mostrados en la Tabla 15 se corresponden con una pareja de altavoces para reproducción de audio en formato estéreo.

Especificaciones

Tal como queda establecido en el apartado 6.1.3 del documento, la prueba de efectividad del modelo de ataque debe ir precedida por una comprobación de la viabilidad de emitir señales no audibles con dispositivos emisores de uso comercial. La máxima frecuencia hasta la cual los fabricantes aseguran una respuesta lineal de los componentes del sistema es 20 kHz en ambos casos.

El objetivo de esta prueba será, por tanto, estudiar el comportamiento de cada sistema emisor cuando las frecuencias de las señales a emitir superan el valor umbral establecido.

Se emplearán diferentes técnicas de modulación con frecuencias de portadora en el rango 20-30 kHz, registrando el resultado en cada caso. Dicho resultado se corresponderá a dos comprobaciones consecutivas e independientes:

- Comprobación de **audibilidad**: En primer lugar, se considerará que el resultado obtenido es positivo si la señal reproducida resulta inaudible para un oyente situado a una distancia aproximada de 1 metro de la fuente de audio.

- Comprobación de **inteligibilidad**: En segundo lugar, se considerará un resultado positivo si la señal reproducida resulta inteligible para un oyente situado a una distancia aproximada de 1 metro de la fuente de audio.

6.2.3.2 Resultados

Comprobación de audibilidad

	DBL Coseno	*DBL Simple*	*BLU estándar*	*BLU Hilbert*
SBS 260	Negativo	Negativo	Negativo	Negativo
Z120	Negativo	Negativo	Negativo	Negativo

Tabla 16. Resultados obtenidos en la prueba de audibilidad para cada dispositivo emisor

Comprobación de inteligibilidad

	DBL Coseno	*DBL Simple*	*BLU estándar*	*BLU Hilbert*
SBS 260	Positivo	Positivo	Negativo	Negativo
Z120	Positivo	Positivo	Negativo	Negativo

Tabla 17 Resultados obtenidos en la prueba de inteligibilidad para cada dispositivo emisor

6.2.3.3 Conclusiones

Ninguna de las técnicas de modulación empleadas genera una señal inaudible en el rango estudiado. La explicación más plausible para esta observación es que la característica no lineal de los altavoces actúa sobre la señal modulada, generando a la salida componentes en frecuencia que no aparecían en la señal original. Esto explicaría que se esté reproduciendo una señal audible cuando teóricamente no debería serlo.

La generalización de este resultado supondría la invalidez del modelo de ataque planteado con un sistema de reproducción de audio con característica lineal no superior a los 20 kHz . La principal implicación de esta afirmación sería la reducción del nivel de riesgo de un ataque mediante señales inaudibles a un sistema asistente de voz: al ser 20 kHz un estándar de facto para el valor umbral superior del rango lineal en el mercado de los reproductores de audio, se reduce la posibilidad de adquirir un altavoz que cumpla con las especificaciones necesarias para el ataque, disminuyendo, por tanto, las probabilidades de que un atacante pudiera llegar a materializar el ataque.

En cuanto a la prueba de inteligibilidad, se obtiene un resultado positivo para las modulaciones de doble banda lateral. Este resultado se tomará como base para la realización de pruebas subsiguientes encaminadas a aumentar la compresión y determinar las implicaciones de la característica no lineal de los altavoces empleados.

6.2.4 PRUEBA DE EFECTIVIDAD DEL ATAQUE

6.2.4.1 Dispositivos empleados y especificaciones de la prueba

Para esta prueba se han utilizado los dispositivos emisores de audio especificados en la Tabla 15, así como los receptores asistentes de voz registrados en la Tabla 5.

- Por un lado, para poder comparar los resultados para distintos asistentes de voz, se empleará como emisor común el dispositivo *SBS 260.*

- Por otro lado, a fin de establecer una comparación entre las prestaciones de distintos altavoces, se empleará como receptor de audio común el dispositivo *Amazon Echo.*

Especificaciones

El objetivo de esta prueba es obtener más información relativa a la característica no lineal de los dispositivos emisores de señal empleados.

En la prueba anterior, se pudo comprobar la reproducción de señales audibles e inteligibles a partir de señales teóricamente no audibles. En este caso, se tratará de comprobar si las

señales reproducidas pueden ser interpretadas por un *software* de reconocimiento de voz. A fin de establecer una comparación entre la información obtenida en esta prueba y los resultados recogidos en el documento explicativo de los *DolphinAttacks*, se empleará el mismo tipo de modulación utilizada en este estudio: doble banda lateral con coseno.

Se emitirá una señal de actuación interpretable por todos los dispositivos un total de 5 veces para cada valor de frecuencia de portadora, el cual variará entre 20 kHz y 30 kHz a intervalos de 100 Hz. Para cada dispositivo, se registrará el rango de frecuencias de portadora para el cual el ataque es *efectivo* (la señal es interpretada correctamente por el dispositivo en, al menos, una ocasión) y el rango considerado como *ideal* (la señal es interpretada correctamente en todos los intentos).

Como información adicional, de manera análoga al estudio de *DolphinAttacks*, se registrará el valor mínimo de índice de modulación (μ) y la distancia máxima para la cual se ha podido conseguir un ataque *ideal*, así como la frecuencia o rango de frecuencias de portadora utilizado en cada caso.

6.2.4.2 Resultados

Comparación entre receptores diferentes:

Dispositivo: Amazon Echo 2ª generación

Rango efectivo (kHz)	*Rango ideal (kHz)*	*Mínimo μ y frecuencia asociada (kHz)*		*Máxima distancia (cm) y frecuencia asociada (kHz)*	
23-28.6	23.6-25	0.2	23.6-23.7	160	24.2

Tabla 18. Resultados obtenidos en la prueba de efectividad del ataque para el receptor Amazon Echo

Dispositivo: Ipad Mini 2

Rango efectivo (kHz)	*Rango ideal (kHz)*	*Mínimo μ y frecuencia asociada (kHz)*		*Máxima distancia y frecuencia asociada (kHz)*	
22.7-25.6	23-24	0.15	23-23.4	170	24

Tabla 19. Resultados obtenidos en la prueba de efectividad del ataque para el receptor Ipad Mini 2

Dispositivo: Samsung Galaxy S21 / Bixby

Rango efectivo (kHz)	*Rango ideal (kHz)*	*Mínimo μ y frecuencia asociada (kHz)*		*Máxima distancia y frecuencia asociada (kHz)*	
23.3-25.3	23.7-23.8	0.5	23.7-23.8	60	23.7

Tabla 20. Resultados obtenidos en la prueba de efectividad del ataque para el receptor Samsung Galaxy S21/ Bixby

Dispositivo: Samsung Galaxy S21 / Google Assistant

Rango efectivo (kHz)	*Rango ideal (kHz)*	*Mínimo μ y frecuencia asociada (kHz)*		*Máxima distancia y frecuencia asociada (kHz)*	
22.5-28.5	22.5-26.4	0.2	23.5	210	26

Tabla 21. Resultados obtenidos en la prueba de efectividad del ataque para el receptor Samsung Galaxy S21/ Google Assistant

Comparación entre emisores diferentes:

Dispositivo: SBS 260

Rango efectivo (kHz)	*Rango ideal (kHz)*	*Mínimo μ y frecuencia asociada (kHz)*		*Máxima distancia (cm) y frecuencia asociada (kHz)*	
23-26.6	23.6-25	0.2	23.6-23.7	130	24.2

Tabla 22. Resultados obtenidos en la prueba de efectividad del ataque para el emisor SBS 260

Dispositivo: Z120

Rango efectivo (kHz)	*Rango ideal (kHz)*	*Mínimo μ y frecuencia asociada (kHz)*		*Máxima distancia (cm) y frecuencia asociada (kHz)*	
23.3-24.1	23.5-23.8	0.2	23.5-23.6	170	24

Tabla 23. Resultados obtenidos en la prueba de efectividad del ataque para el emisor Z120

6.2.4.3 Conclusiones

Se ha comprobado que todos los dispositivos receptores han sido capaces de interpretar las señales emitidas para un cierto rango de frecuencias. **Este resultado demuestra que las réplicas de una señal en banda base, generadas a partir de una respuesta en frecuencia no lineal a una señal modulada en amplitud, pueden ser interpretadas por un sistema asistente de audio.**

Como se puede observar, tanto la comparación entre diferentes asistentes como entre diferentes emisores ofrece resultados variables. No obstante, es difícil aportar una explicación en detalle de la variabilidad de estos resultados sin conocer en profundidad las

características componentes electrónicos empleados por emisor y receptor, con su correspondiente respuesta en frecuencia.

Implicaciones de la respuesta no lineal de los altavoces

La conclusión principal obtenida tras la realización de esta prueba surge a raíz de comparar los datos recogidos en el transcurso de la misma (para el dispositivo *SBS 260*) con los resultados expuestos en el estudio del modelo de ataque de *DolphinAttacks*. Según la información aportada por los investigadores, la respuesta de un dispositivo con asistente de voz tipo *Alexa* al ataque mediante señales no audibles es la siguiente:

Rango efectivo (kHz)	*Rango ideal (kHz)*	*Mínimo μ y frecuencia asociada (kHz)*		*Máxima distancia (cm) y frecuencia asociada (kHz)*	
23-31	24	0.2	24	165	24

Tabla 24. Resultados de prueba del modelo DolphinAttack sobre un dispositivo Amazon Echo. Fuente: (Zhang, y otros, 2017)

Como se puede comprobar, los resultados obtenidos presentan un elevado nivel de coincidencia con los recogidos en esta prueba para el mismo asistente de voz (Tabla 18). En ambos casos, la señal interpretada por los receptores es producto de aplicar una transformación no lineal a una señal modulada en amplitud. La única diferencia es que, en un caso la respuesta no lineal proviene de los componentes del sistema de captación de audio del receptor y, en el otro, es debida directamente a los componentes del emisor. Consecuentemente, podríamos sustituir el esquema del sistema por uno más sencillo, mostrado en la Ilustración 30.

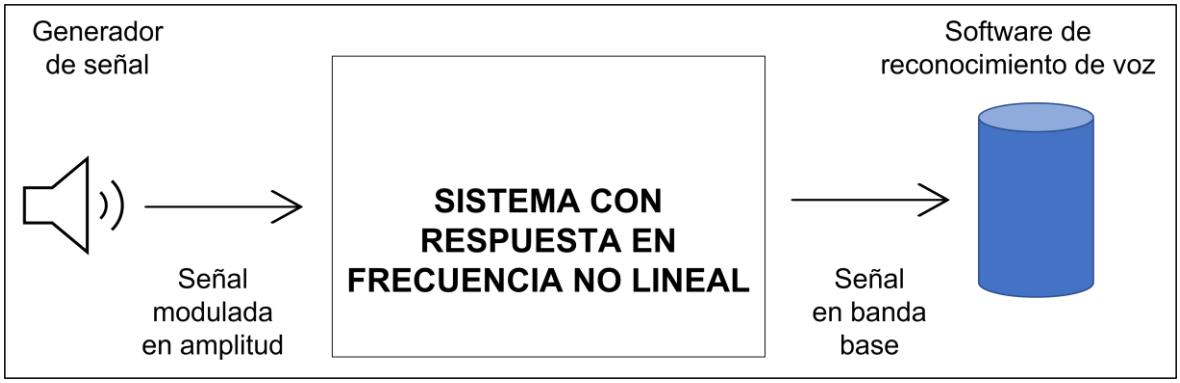


Ilustración 30. Modelo simplificado del aprovechamiento del efecto no linea

Por tanto, se puede concluir que el dispositivo ***SBS 260*** **sirve como modelo de la característica no lineal de un sistema asistente de voz modelo *Alexa*.** Un campo de estudio a desarrollar en trabajos posteriores sería si esta equivalencia se repite para un mayor número de modelos de altavoces y asistentes de voz.

La implicación principal de esta conclusión es que **se posibilita la comprobación de la viabilidad de este modelo de ataque sin necesidad de disponer de un reproductor de audio especialmente diseñado** para emitir frecuencias superiores al umbral del oído humano.

Este factor podría ser aprovechado por un atacante para comprobar la efectividad de un comando o conjunto de comandos específico: bastaría con modular la señal en amplitud dentro del rango de frecuencias adecuado y reproducirla con el sistema de altavoces que mejor modele la respuesta en frecuencia del dispositivo objetivo . Si el asistente interpretase el comando, se podría extrapolar la efectividad de la prueba realizada a un escenario de ataque real.

Del mismo modo, este desarrollo podría ser empleado para probar prototipos de medidas defensivas contra el modelo de ataque estudiado.

6.2.5 PRUEBAS ADICIONALES

Las conclusiones obtenidas en el apartado anterior dejan abierta la posibilidad de obtener una aproximación de la respuesta en frecuencia no lineal del asistente de voz a partir de la señal emitida por los altavoces *SBS 260.*

Por ello, se ha planteado una prueba adicional consistente en capturar la señal emitida por los altavoces para diferentes frecuencias de portadora y representar el espectro en frecuencia de la misma, con el objetivo de comprobar si existe algún tipo de patrón en las frecuencias observables en las señales reproducidas por los altavoces.

Capítulo 7. CONCLUSIONES Y TRABAJOS FUTUROS

La elaboración de este proyecto ha necesitado del desarrollo de un modelo de ataque a dispositivos asistentes de voz mediante señales no audibles. Las prestaciones de dicho modelo, descrito en detalle en el Capítulo 5. Sistema Desarrollado, han sido evaluadas mediante las pruebas recogidas en el Capítulo 6. Análisis de Resultados.

De los resultados obtenidos tras la realización de las pruebas emanan una serie de conclusiones e implicaciones expuestas en los diferentes epígrafes de este Capítulo 6. No obstante, dichas conclusiones serán recogidas en este apartado a fin de contribuir a la organización del documento.

7.1 RESUMEN DE LAS CONCLUSIONES OBTENIDAS

7.1.1 EXISTENCIA DE UN VALOR UMBRAL DE POTENCIA VARIABLE PARA LA CAPTACIÓN DE SEÑALES POR VCS

Los resultados de la prueba explicada en el apartado 6.1.1 demuestran la existencia de un valor mínimo de potencia por debajo del cual las señales captadas por un asistente de voz no son interpretadas.

Aunque estudios anteriores sobre el funcionamiento de los sistemas asistentes de voz ya ponen de manifiesto la existencia de un filtro de potencia en el subsistema de captación de audio de estos dispositivos, durante la fase de experimentación de este proyecto se ha podido comprobar de forma empírica que dicho valor es variable, siempre y cuando el dispositivo cuente con un sistema de activación por comando de voz.

En este sentido, se ha comprobado que la potencia mínima que debe portar una señal de voz para superar los filtros internos de un dispositivo como los estudiados varía en función del tipo de comando que el dispositivo espere recibir. Si el dispositivo se encuentra inactivo y espera, por tanto, recibir un comando de activación, la potencia mínima que necesita dicha

señal para ser susceptible de ser interpretada es mayor que si el dispositivo se encuentra activo.

Es asumible que este hecho se debe a un criterio de diseño de la compañía desarrolladora, probablemente adoptado con el propósito de minimizar los casos de activación accidental del dispositivo. No obstante, esta particularidad también puede suponer una ventaja desde el punto de vista de la seguridad del mismo.

Las técnicas de modulación empleadas para ejecutar la mayoría de modelos de ataque basados en el uso de señales no audibles tienen en común que provocan una disminución de la potencia de la señal original. Además, como también se ha podido comprobar con los resultados de la prueba anteriormente referenciada, las señales emitidas son atenuadas en su propagación hasta el dispositivo. Consecuentemente, cuanto mayor sea el umbral mínimo de potencia del dispositivo, más robusto será frente a modelos de ataque como el desarrollado.

Se puede, por tanto, concluir que un asistente de voz que requiera de un mecanismo de activación por comando de voz será más seguro que un asistente que no cuente con este sistema (frente a un ataque basado en el uso de señales no audibles)

7.1.2 RELACIÓN ENTRE LA ARQUITECTURA DE UN DISPOSITIVO ASISTENTE DE VOZ Y LA POTENCIA DE LAS SEÑALES CAPTADAS

Los resultados de la prueba expuesta en el apartado 6.1.1 ponen de manifiesto una relación entre la estructura interna del dispositivo receptor y la potencia de las señales captadas por este.

Dicha relación ha sido detectada a partir de la aparición datos atípicos en el estudio de la potencia de señal necesaria para activar estos dispositivos a diferentes distancias. Teniendo en cuenta que el sonido es una onda de potencia y este tipo de ondas se atenúan en su propagación por el medio aéreo, lo lógico es pensar que, a menor distancia, menor potencia de señal debería ser requerida para activar el dispositivo.

No obstante, los resultados registrados en el apartado 6.2.1.2 demuestran que esta afirmación puede no cumplirse para algunos sistemas. Esto es debido a que la localización de los micrófonos destinados a la captación de audio en estos artilugios es un factor clave en el proceso de captación de señales de audio.

Dispositivos con una estructura interna determinada (principalmente, aquellos que cuenten con varios micrófonos separados entre sí) pueden tener problemas para captar señales si estas son emitidas a una distancia demasiado reducida de los mismos.

La principal implicación de esta conclusión, dentro del contexto del proyecto, es que un dispositivo con una arquitectura interna como la descrita en la Ilustración 26. será más robusto frente a cualquier modelo que utilice ondas de sonido como vector de ataque, siempre y cuando el emisor se encuentre a una distancia reducida del dispositivo.

7.1.3 USO DE AURICULARES COMO POSIBLE MEDIDA DEFENSIVA

La prueba de desviación en frecuencia, expuesta en el apartado 6.1.2, trata de determinar cuánto puede desviarse una señal en banda base con respecto de la frecuencia central de 0 Hz hasta dejar de ser comprendida por un sistema asistente de voz. De manera equivalente, se podría decir que los resultados de la prueba recogen la frecuencia máxima de portadora que se puede emplear en la modulación de una señal de voz para que la señal modulada siga resultando entendible por el dispositivo.

En el contexto del trabajo, se ha considerado que este valor de frecuencia se encuentra directamente relacionado con la robustez de un VCS frente a un modelo de ataque como el desarrollado.

El éxito del modelo propuesto depende de que la característica no lineal de los componentes del dispositivo objetivo sea equivalente a realizar un proceso de demodulación de la señal de ataque. Incluso si esto ocurre, puede darse la situación de que, al producirse la demodulación, la señal demodulada en banda base no se sitúe exactamente centrada en la frecuencia de 0 Hz, sino que se encuentre desviada un cierto número de hercios. Cuanto

menos tolerante sea un dispositivo a esta situación de deriva frecuencial de la señal, más seguro será frente al modelo de ataque.

La comparación entre los resultados de diferentes asistentes de voz para esta prueba no sugiere diferencias significativas, ni siquiera considerando los dispositivos de últimas generaciones. Sin embargo, se ha podido comprobar cómo el uso de un sistema de auriculares como punto de entrada de audio reduce considerablemente la desviación en frecuencia máxima tolerada por el asistente de voz.

Consecuentemente, se propone el uso de auriculares como posible medida defensiva contra un modelo de ataque basado en señales no audibles, pudiendo ser una hipótesis interesante para ser estudiada en trabajos futuros.

7.1.4 IMPOSIBILIDAD DE REPRODUCIR UN ATAQUE CON SEÑALES NO AUDIBLES HACIENDO USO DE UN ALTAVOZ DE CARÁCTER COMERCIAL

La denominada “prueba de audibilidad”, descrita en el apartado 6.1.3, pone de manifiesto la imposibilidad de reproducir señales no audibles utilizando sistemas de altavoces de carácter comercial.

El apelativo “carácter comercial” se aplica a aquellos sistemas de emisión de audio que se pueden adquirir en los principales centros distribuidores de productos electrónicos. La característica común de este tipo de dispositivos es que no aseguran una respuesta en frecuencia lineal a partir de 20 kHz, siendo este valor un estándar de facto en la industria.

Durante el desarrollo del trabajo ha quedado probado que la respuesta en frecuencia de emisores con estas características impide reproducir señales no audibles, pues el efecto de la no linealidad sobre las señales a reproducir provoca que se generen componentes de frecuencia dentro del rango audible.

La principal implicación de esta conclusión es una reducción del nivel de riesgo de modelos de ataque similares al descrito en el apartado 5.2.

El hecho de que este tipo de ataques requiera de dispositivos emisores especializados en la reproducción de ultrasonidos, los cuales son caros y difíciles de adquirir, va en detrimento de la probabilidad de que un atacante pueda desarrollar el modelo de ataque anteriormente descrito.

7.1.5 POSIBILIDAD DE UTILIZAR ALTAVOCES PARA MODELAR LA RESPUESTA NO LINEAL DE UN ASISTENTE DE VOZ

Tal como muestran los resultados expuestos en el apartado 6.2.3.2, las señales que los altavoces empleados reproducen, cuando se les suministra como entrada una señal modulada por encima de 20 kHz , son entendibles para ciertas frecuencias de portadora.

Se ha podido concluir que la característica no lineal de los componentes electrónicos de estos altavoces genera réplicas de la señal en banda base que sí pueden ser entendidas por un sistema asistente de voz.

Además, al comparar las frecuencias de portadora necesarias para que dichas señales sean entendibles con las frecuencias de portadora necesarias para que funcione el modelo *DolphinAttack* (en el cual se basa este proyecto), se ha observado un elevado rango de coincidencia entre ambos conjuntos de valores para un disposito concreto (*Amazon Echo)*

A partir de este resultado es posible teorizar que, si la misma transformación no lineal que en este caso ocurre en los altavoces del emisor tuviese lugar en los micrófonos del receptor, la señal generada también sería entendible por el dispositivo. En ese caso, los altavoces podrían brindar un método indirecto para comprobar la efectividad del modelo de ataque.

Las conclusiones obtenidas invitar a plantear la hipótesis de que la respuesta en frecuencia no lineal de un sistema emisor de audio se asemeja a la de un dispositivo asistente de voz y, por tanto, podría utilizarse como modelo. La comprobación de la veracidad de esta hipótesis queda planteada como posible campo de estudio de un trabajo posterior.

7.2 TRABAJOS FUTUROS: POSIBLES MEDIDAS DE DEFENSA

7.2.1 IMPLEMENTACIÓN DE MECANISMOS DE AUTENTICACIÓN POR VOZ

La industria tecnológica ha propuesto diferentes contramedidas para proteger los dispositivos de ataques que puedan hacer uso de vulnerabilidades en las herramientas de control por voz. Una de las más importantes y extendidas es la implementación de mecanismos de autenticación basados en voz, los cuales son sistemas de control de acceso biométrico que permiten validar la identidad de un usuario a partir de muestras de su voz (Melin, 2006).

El uso de la voz como elemento distintivo e identificador de una persona es una apuesta que se comenzó a hacer desde el sector bancario. Las cualidades del aparato vocal humano, dependientes de rasgos característicos de la persona como su morfología o costumbres fonéticas, se han demostrado útiles para llevar a cabo labores de reconocimiento e identificación de un individuo (BBVA, 2021).

Aunque difieren en la forma de ser implementados, la base de los algoritmos de identificación de voz es la misma: construir un modelo matemático a partir de muestras de voz que se suministran como entrada al sistema. La forma de construir ese modelo es precisamente lo que diferencia unas implementaciones de otras y donde la industria tecnológica tiene margen de mejora para incrementar la robustez y fiabilidad de estos sistemas.

En el supuesto de integración de uno de estos algoritmos de identificación con un *software* de reconocimiento de voz, habría que modelar inicialmente la voz del usuario que va a gozar de privilegios en el sistema y almacenar las características de ese modelo. En usos posteriores del sistema, bastaría con construir un modelo análogo a partir de las muestras de voz que se reciban a la entrada del sistema y compararlas con el modelo previamente autorizado. Si el resultado de la comparación se encontrase dentro de unos márgenes de error preestablecidos, sería considerado como favorable y, por tanto, se habilitarían todos los permisos de los que gozase un usuario autorizado.

No obstante, este mecanismo de seguridad ya se ha demostrado vulnerable por estudios como el realizado en 2015 por Mukhopadhyay y otros. Este trabajo de investigación demostró que era posible generar señales de voz sintéticas a partir de muestras de voz de una persona tomadas de forma aislada y utilizar estas señales para burlar el funcionamiento de algoritmos de verificación de la identidad de un usuario controlados por voz.

Este modelo de ataque parte de la premisa de que el atacante debe tener acceso a un *set* de muestras de voz de la persona atacada, para poder construir a partir de ellas un modelo matemático, de forma idéntica a como lo haría un algoritmo de identificación por voz. La aportación de estos investigadores se basa en la programación de una red neuronal específicamente entrenada para tomar como entrada los datos de dicho y modelo matemático y sintetizar a partir de ellos muestras de audio que simulen las características de la voz a la que el modelo daba forma, teniendo la posibilidad de emitir frases o expresiones a la elección del atacante.

El hecho de que el atacante deba contar con una selección de muestras de voz de su objetivo lo suficientemente grande como para entrenar debidamente a la red neuronal supone una limitación al modelo de ataque, pero al mismo tiempo no deja de ser un escenario plausible, ya que en la actualidad se dispone de dispositivos de grabación lo suficientemente sofisticados como para llevar a cabo esta tarea.

Capítulo 8. BIBLIOGRAFÍA

Advsolned. (28 de Abril de 2020). *advsolned.com*. Obtenido de advsolned.com: https://www.advsolned.com/difference-between-iir-and-fir-filters-a-practical-design-guide/

Allen, J. (4 de Octubre de 2021). *techradar.com*. Obtenido de techradar.com: https://www.techradar.com/news/siri-10-year-anniversary

Avisoft Bioacustics. (s.f.). *http://www.avisoft.com/*. Obtenido de http://www.avisoft.com/: http://www.avisoft.com/price-list-ordering-information/

BBVA. (23 de Abril de 2021). *bbva.com*. Obtenido de bbva.com: https://www.bbva.com/es/biometria-de-voz-la-huella-vocal-sera-el-gran-aliado-de-la-banca-online/

Brüel & Kjær. (s.f.). *bksv.com*. Obtenido de https://www.bksv.com/es/transducers

Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., . . . Zhou, W. (2016). *Hidden Voice Commands.* USENIX Security Symposium.

Computerworld. (14 de Junio de 2019). Recuperado el 10 de 11 de 2021, de https://www.computerworld.es/convergencia/el-mercado-de-dispositivos-inteligentes-para-el-hogar-movera-75000-millones-de-dolares-en-2025

Diao, W., Liu, X., Zhou, Z., & Zhang, K. (2014). *Your Voice Assistant is Mine: How to Abuse Speakers to Steal Information and Control Your Phone.* Hong Kong: Department of Information Engineering. The Chinese University of Hong Kong.

Esteves, C. K. (2015). *IEMI Threats for Information Security: Remote Command Injection.* Paris.

GlobalInfoResearch. (2020). *Global Speech Recognition Market 2020 by Company, Regions, Type and Application, Forecast to 2025.*

González, N. (10 de Junio de 2014). *gadgethacks.com*. Obtenido de gadgethacks.com: https://ios.gadgethacks.com/how-to/siri-exploited-again-bypass-lock-screen-ios-8-protect-yourself-0157749/

Kaberpanthi, N., & Datar, A. (2014). Speaker Independent Speech Recognition using MFCC. *International Journal of Computer Applications*, Volume 95– No.26.

Kasmi, C., & Esteves, J. L. (2015). IEMI Threats for Information Security: Remote Command Injection. *IEEE TRANSACTIONS ON ELECTROMAGNETIC COMPATIBILITY*, VOL. 57, NO. 6.

Melin, H. (2006). *Automatic speaker verification on site and by telephone: methods, applications and assessment.* Stockholm, Sweden.

Purves, D., Augustine, G. J., Fitzpatrick, D., Hall, W. C., LaMantia, A.-S., & White, L. E. (2012). *Neuroscience.* Sunderland, Massachusetts: Sinauer Associates.

Quintanilla, N. (5 de Enero de 2018). *https://www.elperiodico.com/*. Obtenido de https://www.elperiodico.com/: https://www.elperiodico.com/es/sociedad/20180105/fallo-seguridad-intel-moviliza-empresas-6533921

Rose, B. (21 de Febrero de 2019). *CUI Devices*. Obtenido de https://www.cuidevices.com/blog/comparing-mems-and-electret-condenser-microphones

Roy, N., Hassanieh, H., & Choudhury, R. R. (2017). *BackDoor: Making Microphones Hear Inaudible Sounds.*

Texas Instruments. (Octubre de 2019). *Ti.com.* Obtenido de Ti.com: https://www.ti.com/lit/ds/symlink/tlv320adc5140.pdf?ts=1656151818468&ref_url=https%253A%252F%252Fwww.google.com%252F

Vaidya, T., Zhang, Y., Sherr, M., & Shields, C. (2015). Cocaine Noodles: Exploiting the Gap between Human and Machine Speech Recognition. *In Proceedings of the 9th USENIX Conference on Offensive Technologies (WOOT'15). USENIX Association, USA, 16.*

Wilczek, J. (28 de Noviembre de 2019). *https://thewolfsound.com/*. Obtenido de https://thewolfsound.com/: https://thewolfsound.com/what-is-aliasing-what-causes-it-how-to-avoid-it/

Yan, Q., Liu, K., Zhou, Q., Guo, H., & Zhang, N. (2020). *SurfingAttack: Interactive Hidden Attack on Voice Assistants Using Ultrasonic Guided Waves.* Conference: Network and Distributed System Security Symposium.

Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., & Xu, W. (2017). DolphinAttack: Inaudible Voice Commands. *ACM CCS.* Dallas, Tx.

ANEXO I: ALINEACIÓN DEL PROYECTO CON LOS ODS

Los Objetivos de Desarrollo Sostenible (ODS) son una serie de metas pactadas por los Estados Miembros de las Naciones Unidas en el contexto de la Agenda 2030. Cada uno de estos objetivos (diecisiete en total) abarca una serie de propósitos y líneas de actuación que persiguen, dentro de un contexto determinado, la consecución de una serie de transformaciones a nivel mundial encaminadas a asegurar un desarrollo social próspero y sostenible durante los próximos años.

Cada uno de estos objetivos hace referencia a una línea de trabajo concreta por la cual se pueden lograr dichas transformaciones, como podría ser la erradicación de la pobreza o el establecimiento de unos patrones de consumo responsables.

Se puede considerar la alineación del proyecto realizado con los ODS desde dos perspectivas diferentes: desde la perspectiva de una relación directa con el objetivo número 9, *Industria, Innovación e Infraestructuras*, y desde la perspectiva de una relación transversal con el resto de objetivos.

Relación directa con el objetivo 9: *Industria, Innovación e Infraestructuras*

Dentro de las metas establecidas en el marco de este objetivo, se indica la necesidad de *"Aumentar la investigación científica y mejorar la capacidad tecnológica de los sectores industriales de todos los países"*

Precisamente, mediante el estudio de las vulnerabilidades asociadas a los VCS se está contribuyendo al desarrollo de la industria tecnológica, centrada en ese sector concreto. Si una vulnerabilidad de un dispositivo electrónico es descubierta y explotada, las consecuencias negativas que afrontan los fabricantes son realmente severas.

Recientemente, se han dado casos de empresas que experimentaban considerables pérdidas económicas tras descubrirse un fallo en alguno de los productos que comercializa. Cualquier

efecto negativo contra la industria tecnológica iría directamente en contra del progreso y el desarrollo sostenible que se persigue con este objetivo.

Por ello, estudios que persigan la caracterización de vulnerabilidades en dispositivos electrónicos y traten de proponer medidas de defensa ante ellos van de la mano con la meta fijada por este punto de los ODS.

Relación transversal con la totalidad de los ODS

Otra forma de entender la relación de este proyecto con la consecución de los ODS consiste en valorar el carácter transversal de la industria tecnológica en la sociedad actual. El elevado peso que la tecnología ha adquirido en nuestra sociedad en las últimas dos décadas es responsable de que, a día de hoy, la evolución de la industria tecnológica condicione el desarrollo del resto de sectores de la sociedad.

Por ello, se puede afirmar que, no solo este, sino cualquier otro proyecto de carácter técnico que busque contribuir al desarrollo de la industria tecnológica se encuentra alineado con los objetivos planteados en los ODS.

ANEXO II. HERRAMIENTA SOFTWARE PARA MODULACIÓN DE SEÑALES

El siguiente Anexo sirve de muestra de imágenes y código fuente de la herramienta software desarrollada en Matlab y empleada para ejecutar las diferentes técnicas de modulación que se referencian en el apartado 5.2.2.

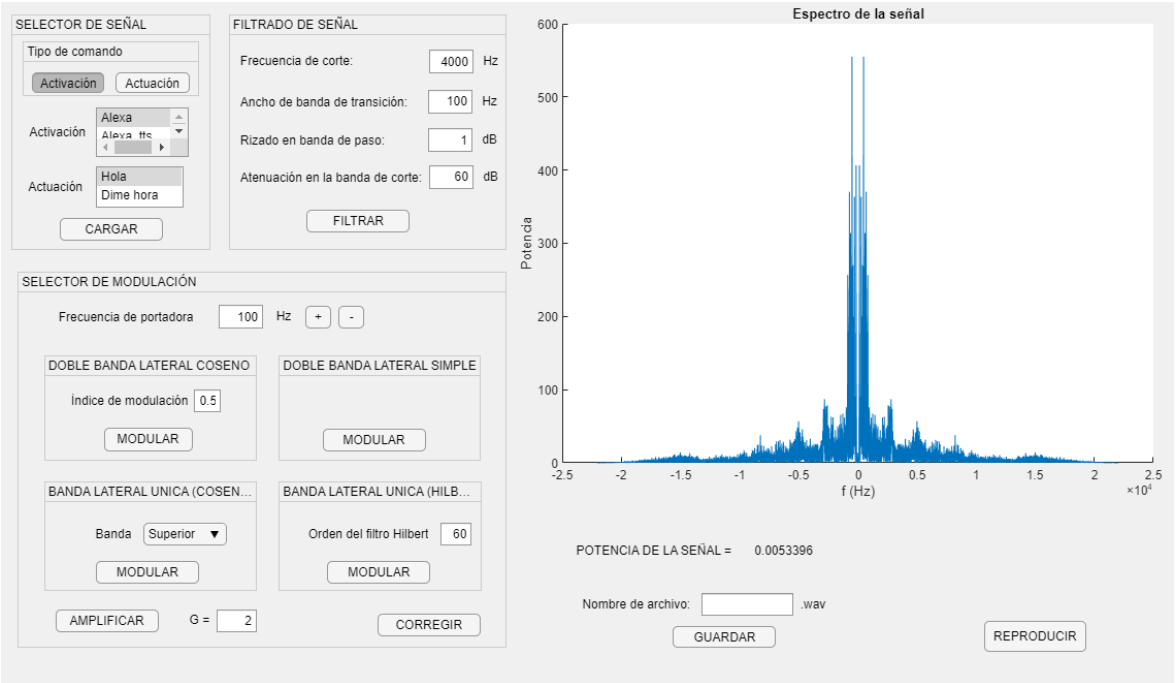


Ilustración 31. Vista general de la aplicación

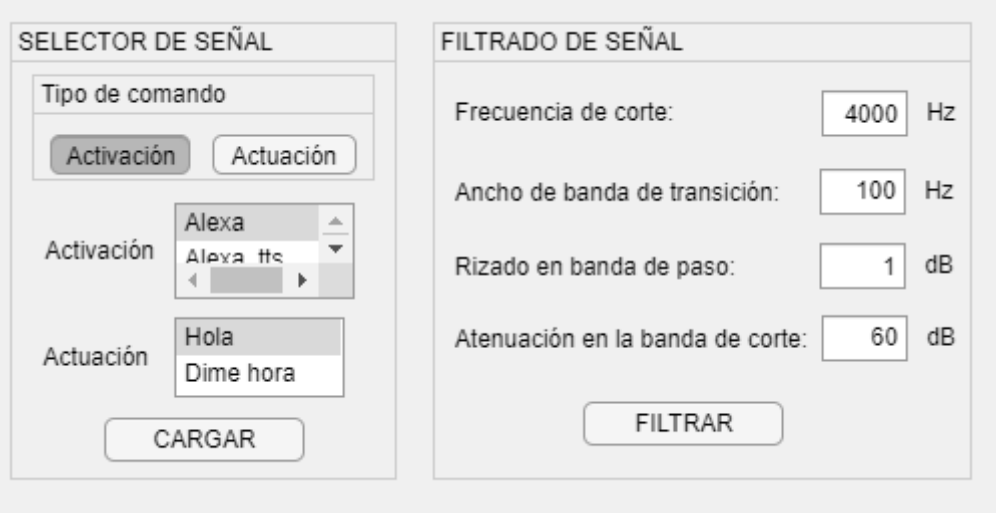


Ilustración 32. Panel de selección de señal

Ilustración 33. Panel de selección de modulación

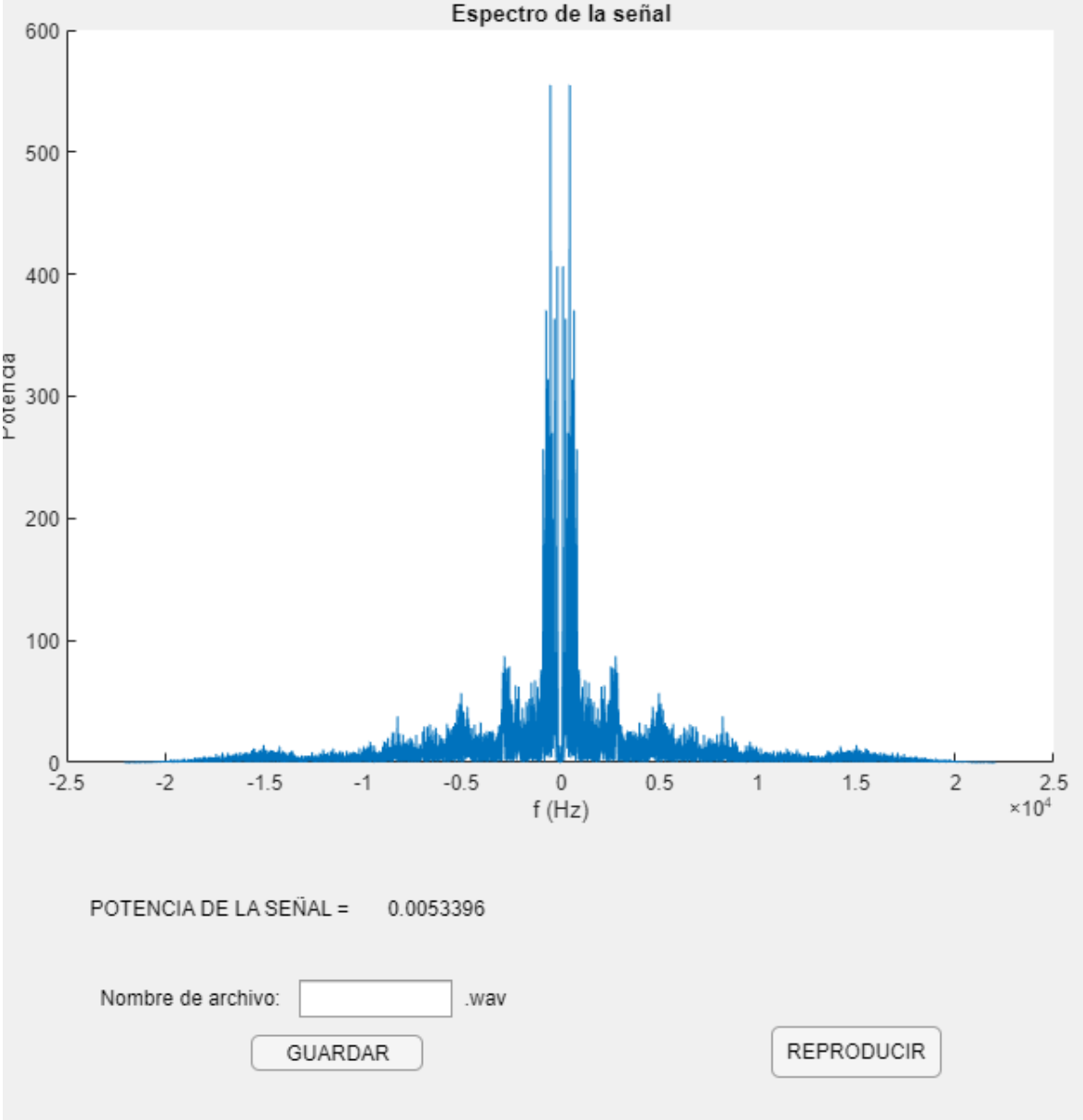


Ilustración 34. Panel de muestra, almacenaje y reproducción de señales

Código fuente de la aplicación:

```
classdef modulatorApp_v2 < matlab.apps.AppBase


% Properties that correspond to app components
properties (Access = public)
UIFigure matlab.ui.Figure
PotenciaLabel matlab.ui.control.Label
PotenciaTextLabel matlab.ui.control.Label
GuardadoLabel matlab.ui.control.Label
wavLabel matlab.ui.control.Label
GUARDARButton matlab.ui.control.Button
FileNameEditField matlab.ui.control.EditField
NombredearchivoLabel matlab.ui.control.Label
SignalFilterPanel matlab.ui.container.Panel
AtCorteLabel matlab.ui.control.Label
RipPasoLabel matlab.ui.control.Label
AtCorteEditField matlab.ui.control.NumericEditField
dBLabel_2 matlab.ui.control.Label
FcorteEditField matlab.ui.control.NumericEditField
HzEditFieldLabel_3 matlab.ui.control.Label
RipPasoEditField matlab.ui.control.NumericEditField
dBLabel matlab.ui.control.Label
BwTransLabel matlab.ui.control.Label
FcorteLabel matlab.ui.control.Label
BwTransEditField matlab.ui.control.NumericEditField
HzEditFieldLabel matlab.ui.control.Label
FILTRARButton matlab.ui.control.Button
SELECTORDEMODULACINPanel matlab.ui.container.Panel
CORREGIRButton matlab.ui.control.Button
GEditField matlab.ui.control.NumericEditField
GLabel matlab.ui.control.Label
AMPLIFICARButton matlab.ui.control.Button
RestarButtton matlab.ui.control.Button
SumarButton matlab.ui.control.Button
FPortadoraLabel matlab.ui.control.Label
FPortadoraEditField matlab.ui.control.NumericEditField
dBLabel_3 matlab.ui.control.Label
BANDALATERALUNICAHILBERTPanel matlab.ui.container.Panel
OrdendelfiltroHilbertEditField matlab.ui.control.NumericEditField
OrdendelfiltroHilbertEditFieldLabel matlab.ui.control.Label
MODULAR_bluhilbert_Button matlab.ui.control.Button
BANDALATERALUNICACOSENOPanel matlab.ui.container.Panel
BandaDropDown matlab.ui.control.DropDown
BandaDropDownLabel matlab.ui.control.Label
MODULAR_blucoseno_Button matlab.ui.control.Button
DOBLEBANDALATERALSIMPLEPanel matlab.ui.container.Panel
MODULAR_exp_Button matlab.ui.control.Button
DOBLEBANDALATERALCOSENOPanel matlab.ui.container.Panel
```

```
ModIndexEditField matlab.ui.control.NumericEditField
ndicedemodulacinEditFieldLabel matlab.ui.control.Label
MODULAR_dbl_Button matlab.ui.control.Button
SignalSelectorPanel matlab.ui.container.Panel
CommandTypeButtonGroup matlab.ui.container.ButtonGroup
TipoActuacionButton matlab.ui.control.ToggleButton
TipoActivacionButton matlab.ui.control.ToggleButton
ActuacionListBox matlab.ui.control.ListBox
ActuacinListBoxLabel matlab.ui.control.Label
ActivacionListBox matlab.ui.control.ListBox
ActivacinListBoxLabel matlab.ui.control.Label
CARGARButton matlab.ui.control.Button
REPRODUCIRButton matlab.ui.control.Button
UIAxes matlab.ui.control.UIAxes
end


% Callbacks that handle component events
methods (Access = private)


% Button pushed function: CARGARButton
function CARGARButtonPushed(app, event)
global Fs;
global x_t;
global t;
global rep;
if app.TipoActivacionButton.Value
command = app.ActivacionListBox.Value;
else
command = app.ActuacionListBox.Value;
end
if strcmp('Alexa',command)
[x_t,Fs]= audioread('audios_og/Alexa2.wav');
Ts= 1/Fs;
t= linspace(0,(length(x_t)-1)*Ts,length(x_t))';
elseif strcmp('Siri',command)
[x_t,Fs]= audioread('audios_og/OyeSiri.wav');
Ts= 1/Fs;
t= linspace(0,(length(x_t)-1)*Ts,length(x_t))';
elseif strcmp('Hola',command)
[x_t,Fs]= audioread('audios_og/hola.wav');
Ts= 1/Fs;
t= linspace(0,(length(x_t)-1)*Ts,length(x_t))';
elseif strcmp('Alexa_tts',command)
[x_t,Fs]= audioread('audios_og/Alexa_tts.wav');
Ts= 1/Fs;
t= linspace(0,(length(x_t)-1)*Ts,length(x_t))';
end
```

```
y_t = x_t;


n=length(y_t);
f=(-n/2:n/2-1)*(Fs/n);
Y_f = fftshift(fft(y_t));
plot(app.UIAxes,f,abs(Y_f));


app.PotenciaLabel.Text = num2str(get_potencia(y_t));


rep= y_t;
end


% Button pushed function: FILTRARButton
function FILTRARButtonPushed(app, event)
global Fs;
global x_t;
global rep;


fc = app.FcorteEditField.Value;


BwTrans = app.BwTransEditField.Value;
Fpass = fc-BwTrans/2;
Fstop = fc+BwTrans/2;
Ap = app.RipPasoEditField.Value;


Ast = app.AtCorteEditField.Value;




d = get_FIR_filter(Fpass,Fstop,Ap,Ast,Fs,0);




D = round(mean(grpdelay(d)));


y_t = filter(d,[x_t;zeros(D,1)]);
y_t= y_t(D+1:end);


n=length(y_t);
f=(-n/2:n/2-1)*(Fs/n);
```

```
Y_f = fftshift(fft(y_t));
plot(app.UIAxes,f,abs(Y_f));


app.PotenciaLabel.Text = num2str(get_potencia(y_t));
x_t=y_t;
rep= y_t;
end


% Button pushed function: MODULAR_dbl_Button
function MODULAR_dbl_ButtonPushed(app, event)
global Fs;
global x_t;
global t;
global rep;


Fw= app.FPortadoraEditField.Value;


mu = app.ModIndexEditField.Value;


y_t = mod_am_cos(x_t,mu,Fw,t,1);


n=length(y_t);
f=(-n/2:n/2-1)*(Fs/n);
Y_f = fftshift(fft(y_t));
plot(app.UIAxes,f,abs(Y_f));
app.UIAxes.YLim=[0 600];


app.PotenciaLabel.Text = num2str(get_potencia(y_t));


rep= y_t;


app.GuardadoLabel.Text=''
end


% Button pushed function: MODULAR_exp_Button
function MODULAR_exp_ButtonPushed(app, event)
global Fs;
global x_t;
global t;
global rep;
```

```
Fw= app.FPortadoraEditField.Value;


y_t = mod_am_DBL_simple(x_t,Fw,t);


n=length(y_t);
f=(-n/2:n/2-1)*(Fs/n);
Y_f = fftshift(fft(y_t));
plot(app.UIAxes,f,abs(Y_f));


app.PotenciaLabel.Text = num2str(get_potencia(y_t));


rep= y_t;


app.GuardadoLabel.Text=''
end


% Button pushed function: MODULAR_blucoseno_Button
function MODULAR_blucoseno_ButtonPushed(app, event)
global Fs;
global x_t;
global t;
global rep;


Fw= app.FPortadoraEditField.Value;
if strcmp(app.BandaDropDown.Value,'Superior')
OPT = 1;
else
OPT = 0;
end
y_t= mod_am_BLU_coseno(x_t,Fw, OPT,Fs,t);


n=length(y_t);
f=(-n/2:n/2-1)*(Fs/n);
Y_f = fftshift(fft(y_t));
plot(app.UIAxes,f,abs(Y_f));


app.PotenciaLabel.Text = num2str(get_potencia(y_t));


rep= y_t;
```

```
app.GuardadoLabel.Text=''
end


% Button pushed function: MODULAR_bluhilbert_Button
function MODULAR_bluhilbert_ButtonPushed(app, event)
global Fs;
global x_t;
global t;
global rep;


Fw= app.FPortadoraEditField.Value;


fo= app.OrdendelfiltroHilbertEditField.Value;


y_t = mod_am_BLU_hilbert(x_t,Fw,fo,Fs,t);
n=length(y_t);
f=(-n/2:n/2-1)*(Fs/n);
Y_f = fftshift(fft(y_t));
plot(app.UIAxes,f,abs(Y_f));


app.PotenciaLabel.Text = num2str(get_potencia(y_t));


rep= y_t;


app.GuardadoLabel.Text=''
end


% Button pushed function: REPRODUCIRButton
function REPRODUCIRButtonPushed(app, event)
global Fs;
global x_t;
global rep;
try
sound(rep,Fs);
catch
sound(real(rep),Fs);
end
end


% Button pushed function: AMPLIFICARButton
```

```
function AMPLIFICARButtonPushed(app, event)
global Fs;
global x_t;
global rep;


y_t = rep;


G = app.GEditField.Value;


y_t= G*y_t;


n=length(y_t);
f=(-n/2:n/2-1)*(Fs/n);
Y_f = fftshift(fft(y_t));
plot(app.UIAxes,f,abs(Y_f));


app.PotenciaLabel.Text = num2str(get_potencia(y_t));


rep = y_t;


app.GuardadoLabel.Text=''
end


% Button pushed function: SumarButton
function SumarButtonPushed(app, event)
app.FPortadoraEditField.Value = app.FPortadoraEditField.Value + 100;
end


% Button pushed function: RestarButtton
function RestarButttonButtonPushed(app, event)
app.FPortadoraEditField.Value = app.FPortadoraEditField.Value - 100;
end


% Button pushed function: GUARDARButton
function GUARDARButtonPushed(app, event)
global Fs;
global x_t;
global rep;
name = app.FileNameEditField.Value;
filename= strcat('audios_mod/',name,'.wav');
audiowrite(filename,rep,Fs);
```

```
app.FileNameEditField.Value = '';
app.GuardadoLabel.Text='Audio Guardado'


end


% Button pushed function: CORREGIRButton
function CORREGIRButtonPushed(app, event)
global Fs;
global x_t;
global rep;



Fw= app.FPortadoraEditField.Value;


bpFilt = designfilt('bandstopfir','FilterOrder',10000, ...
'CutoffFrequency1',Fw-10,'CutoffFrequency2',Fw+10, ...
'SampleRate',Fs);



y_t= filter(bpFilt,rep);


n=length(y_t);
f=(-n/2:n/2-1)*(Fs/n);
Y_f = fftshift(fft(y_t));
plot(app.UIAxes,f,abs(Y_f));


app.PotenciaLabel.Text = num2str(get_potencia(y_t));


rep=y_t;



end


% Callback function
function ATENUARButtonPushed(app, event)
```

```
end
end


% Component initialization
methods (Access = private)


% Create UIFigure and components
function createComponents(app)


% Create UIFigure and hide until all components are created
app.UIFigure = uifigure('Visible', 'off');
app.UIFigure.Position = [100 100 1352 851];
app.UIFigure.Name = 'MATLAB App';


% Create UIAxes
app.UIAxes = uiaxes(app.UIFigure);
title(app.UIAxes, 'Espectro de la señal')
xlabel(app.UIAxes, 'f (Hz)')
ylabel(app.UIAxes, 'Potencia')
zlabel(app.UIAxes, 'Z')
app.UIAxes.Position = [573 231 747 517];


% Create REPRODUCIRButton
app.REPRODUCIRButton = uibutton(app.UIFigure, 'push');
app.REPRODUCIRButton.ButtonPushedFcn = createCallbackFcn(app,
@REPRODUCIRButtonPushed, true);
app.REPRODUCIRButton.Position = [1131 39 99 30];
app.REPRODUCIRButton.Text = 'REPRODUCIR';


% Create SignalSelectorPanel
app.SignalSelectorPanel = uipanel(app.UIFigure);
app.SignalSelectorPanel.Title = 'SELECTOR DE SEÑAL';
app.SignalSelectorPanel.Position = [20 577 221 255];


% Create CARGARButton
app.CARGARButton = uibutton(app.SignalSelectorPanel, 'push');
app.CARGARButton.ButtonPushedFcn = createCallbackFcn(app, @CARGARButtonPushed,
true);
app.CARGARButton.Position = [59 10 100 22];
app.CARGARButton.Text = 'CARGAR';


% Create ActivacinListBoxLabel
```

```
app.ActivacinListBoxLabel = uilabel(app.SignalSelectorPanel);
app.ActivacinListBoxLabel.HorizontalAlignment = 'right';
app.ActivacinListBoxLabel.Position = [14 105 61 22];
app.ActivacinListBoxLabel.Text = 'Activación';


% Create ActivacionListBox
app.ActivacionListBox = uilistbox(app.SignalSelectorPanel);
app.ActivacionListBox.Items = {'Alexa', 'Alexa_tts', 'Siri'};
app.ActivacionListBox.Position = [90 92 96 48];
app.ActivacionListBox.Value = 'Alexa';


% Create ActuacinListBoxLabel
app.ActuacinListBoxLabel = uilabel(app.SignalSelectorPanel);
app.ActuacinListBoxLabel.HorizontalAlignment = 'right';
app.ActuacinListBoxLabel.Position = [14 52 58 22];
app.ActuacinListBoxLabel.Text = 'Actuación';


% Create ActuacionListBox
app.ActuacionListBox = uilistbox(app.SignalSelectorPanel);
app.ActuacionListBox.Items = {'Hola', 'Dime hora'};
app.ActuacionListBox.Position = [90 43 90 40];
app.ActuacionListBox.Value = 'Hola';


% Create CommandTypeButtonGroup
app.CommandTypeButtonGroup = uibuttongroup(app.SignalSelectorPanel);
app.CommandTypeButtonGroup.Title = 'Tipo de comando';
app.CommandTypeButtonGroup.Position = [19 159 180 58];


% Create TipoActivacionButton
app.TipoActivacionButton = uitogglebutton(app.CommandTypeButtonGroup);
app.TipoActivacionButton.Text = 'Activación';
app.TipoActivacionButton.Position = [9 3 70 22];
app.TipoActivacionButton.Value = true;


% Create TipoActuacionButton
app.TipoActuacionButton = uitogglebutton(app.CommandTypeButtonGroup);
app.TipoActuacionButton.Text = 'Actuación';
app.TipoActuacionButton.Position = [98 3 72 22];


% Create SELECTORDEMODULACINPanel
app.SELECTORDEMODULACINPanel = uipanel(app.UIFigure);
app.SELECTORDEMODULACINPanel.Title = 'SELECTOR DE MODULACIÓN';
app.SELECTORDEMODULACINPanel.Position = [23 44 526 498];
```

```
% Create DOBLEBANDALATERALCOSENOPanel
app.DOBLEBANDALATERALCOSENOPanel = uipanel(app.SELECTORDEMODULACINPanel);
app.DOBLEBANDALATERALCOSENOPanel.Title = 'DOBLE BANDA LATERAL COSENO';
app.DOBLEBANDALATERALCOSENOPanel.Position = [32 255 223 130];


% Create MODULAR_dbl_Button
app.MODULAR_dbl_Button = uibutton(app.DOBLEBANDALATERALCOSENOPanel, 'push');
app.MODULAR_dbl_Button.ButtonPushedFcn = createCallbackFcn(app,
@MODULAR_dbl_ButtonPushed, true);
app.MODULAR_dbl_Button.Position = [64 9 86 23];
app.MODULAR_dbl_Button.Text = 'MODULAR';


% Create ndicedemodulacinEditFieldLabel
app.ndicedemodulacinEditFieldLabel = uilabel(app.DOBLEBANDALATERALCOSENOPanel);
app.ndicedemodulacinEditFieldLabel.HorizontalAlignment = 'right';
app.ndicedemodulacinEditFieldLabel.Position = [27 61 119 22];
app.ndicedemodulacinEditFieldLabel.Text = 'Índice de modulación';


% Create ModIndexEditField
app.ModIndexEditField = uieditfield(app.DOBLEBANDALATERALCOSENOPanel, 'numeric');
app.ModIndexEditField.Position = [152 61 26 22];
app.ModIndexEditField.Value = 0.5;


% Create DOBLEBANDALATERALSIMPLEPanel
app.DOBLEBANDALATERALSIMPLEPanel = uipanel(app.SELECTORDEMODULACINPanel);
app.DOBLEBANDALATERALSIMPLEPanel.Title = 'DOBLE BANDA LATERAL SIMPLE';
app.DOBLEBANDALATERALSIMPLEPanel.Position = [285 255 207 130];


% Create MODULAR_exp_Button
app.MODULAR_exp_Button = uibutton(app.DOBLEBANDALATERALSIMPLEPanel, 'push');
app.MODULAR_exp_Button.ButtonPushedFcn = createCallbackFcn(app,
@MODULAR_exp_ButtonPushed, true);
app.MODULAR_exp_Button.Position = [47 9 100 22];
app.MODULAR_exp_Button.Text = 'MODULAR';


% Create BANDALATERALUNICACOSENOPanel
app.BANDALATERALUNICACOSENOPanel = uipanel(app.SELECTORDEMODULACINPanel);
app.BANDALATERALUNICACOSENOPanel.Title = 'BANDA LATERAL UNICA (COSENO)';
app.BANDALATERALUNICACOSENOPanel.Position = [32 78 223 141];


% Create MODULAR_blucoseno_Button
```

```
app.MODULAR_blucoseno_Button = uibutton(app.BANDALATERALUNICACOSENOPanel,
'push');
app.MODULAR_blucoseno_Button.ButtonPushedFcn = createCallbackFcn(app,
@MODULAR_blucoseno_ButtonPushed, true);
app.MODULAR_blucoseno_Button.Position = [58 7 100 22];
app.MODULAR_blucoseno_Button.Text = 'MODULAR';


% Create BandaDropDownLabel
app.BandaDropDownLabel = uilabel(app.BANDALATERALUNICACOSENOPanel);
app.BandaDropDownLabel.HorizontalAlignment = 'right';
app.BandaDropDownLabel.Position = [24 59 65 22];
app.BandaDropDownLabel.Text = 'Banda';


% Create BandaDropDown
app.BandaDropDown = uidropdown(app.BANDALATERALUNICACOSENOPanel);
app.BandaDropDown.Items = {'Superior', 'Inferior'};
app.BandaDropDown.Position = [103 59 84 22];
app.BandaDropDown.Value = 'Superior';


% Create BANDALATERALUNICAHILBERTPanel
app.BANDALATERALUNICAHILBERTPanel = uipanel(app.SELECTORDEMODULACINPanel);
app.BANDALATERALUNICAHILBERTPanel.Title = 'BANDA LATERAL UNICA (HILBERT)';
app.BANDALATERALUNICAHILBERTPanel.Position = [285 78 208 141];


% Create MODULAR_bluhilbert_Button
app.MODULAR_bluhilbert_Button = uibutton(app.BANDALATERALUNICAHILBERTPanel,
'push');
app.MODULAR_bluhilbert_Button.ButtonPushedFcn = createCallbackFcn(app,
@MODULAR_bluhilbert_ButtonPushed, true);
app.MODULAR_bluhilbert_Button.Position = [47 7 100 22];
app.MODULAR_bluhilbert_Button.Text = 'MODULAR';


% Create OrdendelfiltroHilbertEditFieldLabel
app.OrdendelfiltroHilbertEditFieldLabel =
uilabel(app.BANDALATERALUNICAHILBERTPanel);
app.OrdendelfiltroHilbertEditFieldLabel.HorizontalAlignment = 'right';
app.OrdendelfiltroHilbertEditFieldLabel.Position = [1 59 144 22];
app.OrdendelfiltroHilbertEditFieldLabel.Text = 'Orden del filtro Hilbert';


% Create OrdendelfiltroHilbertEditField
app.OrdendelfiltroHilbertEditField =
uieditfield(app.BANDALATERALUNICAHILBERTPanel, 'numeric');
app.OrdendelfiltroHilbertEditField.Position = [159 59 30 22];
app.OrdendelfiltroHilbertEditField.Value = 60;
```

```
% Create dBLabel_3
app.dBLabel_3 = uilabel(app.SELECTORDEMODULACINPanel);
app.dBLabel_3.HorizontalAlignment = 'right';
app.dBLabel_3.Position = [202 434 83 22];
app.dBLabel_3.Text = 'Hz';


% Create FPortadoraEditField
app.FPortadoraEditField = uieditfield(app.SELECTORDEMODULACINPanel, 'numeric');
app.FPortadoraEditField.Position = [218 433 43 23];
app.FPortadoraEditField.Value = 100;


% Create FPortadoraLabel
app.FPortadoraLabel = uilabel(app.SELECTORDEMODULACINPanel);
app.FPortadoraLabel.Position = [55 433 136 22];
app.FPortadoraLabel.Text = 'Frecuencia de portadora';


% Create SumarButton
app.SumarButton = uibutton(app.SELECTORDEMODULACINPanel, 'push');
app.SumarButton.ButtonPushedFcn = createCallbackFcn(app, @SumarButtonPushed,
true);
app.SumarButton.Position = [308 433 25 22];
app.SumarButton.Text = '+';


% Create RestarButtton
app.RestarButttton = uibutton(app.SELECTORDEMODULACINPanel, 'push');
app.RestarButttton.ButtonPushedFcn = createCallbackFcn(app,
@RestarButttonButtonPushed, true);
app.RestarButttton.Position = [341 433 25 22];
app.RestarButttton.Text = '-';


% Create AMPLIFICARButton
app.AMPLIFICARButton = uibutton(app.SELECTORDEMODULACINPanel, 'push');
app.AMPLIFICARButton.ButtonPushedFcn = createCallbackFcn(app,
@AMPLIFICARButtonPushed, true);
app.AMPLIFICARButton.Position = [49 25 100 22];
app.AMPLIFICARButton.Text = 'AMPLIFICAR';


% Create GLabel
app.GLabel = uilabel(app.SELECTORDEMODULACINPanel);
app.GLabel.HorizontalAlignment = 'right';
app.GLabel.Position = [182 25 25 22];
app.GLabel.Text = 'G =';
```

```
% Create GEditField
app.GEditField = uieditfield(app.SELECTORDEMODULACINPanel, 'numeric');
app.GEditField.Position = [216 25 39 22];
app.GEditField.Value = 2;


% Create CORREGIRButton
app.CORREGIRButton = uibutton(app.SELECTORDEMODULACINPanel, 'push');
app.CORREGIRButton.ButtonPushedFcn = createCallbackFcn(app,
@CORREGIRButtonPushed, true);
app.CORREGIRButton.Position = [394 17 100 22];
app.CORREGIRButton.Text = 'CORREGIR';


% Create SignalFilterPanel
app.SignalFilterPanel = uipanel(app.UIFigure);
app.SignalFilterPanel.Title = 'FILTRADO DE SEÑAL';
app.SignalFilterPanel.Position = [278 577 271 255];


% Create FILTRARButton
app.FILTRARButton = uibutton(app.SignalFilterPanel, 'push');
app.FILTRARButton.ButtonPushedFcn = createCallbackFcn(app, @FILTRARButtonPushed,
true);
app.FILTRARButton.Position = [77 22 100 22];
app.FILTRARButton.Text = 'FILTRAR';


% Create HzEditFieldLabel
app.HzEditFieldLabel = uilabel(app.SignalFilterPanel);
app.HzEditFieldLabel.HorizontalAlignment = 'right';
app.HzEditFieldLabel.Position = [179 152 83 22];
app.HzEditFieldLabel.Text = 'Hz';


% Create BwTransEditField
app.BwTransEditField = uieditfield(app.SignalFilterPanel, 'numeric');
app.BwTransEditField.Position = [195 151 43 23];
app.BwTransEditField.Value = 100;


% Create FcorteLabel
app.FcorteLabel = uilabel(app.SignalFilterPanel);
app.FcorteLabel.Position = [13 195 115 22];
app.FcorteLabel.Text = 'Frecuencia de corte:';


% Create BwTransLabel
```

```
app.BwTransLabel = uilabel(app.SignalFilterPanel);
app.BwTransLabel.Position = [13 151 168 22];
app.BwTransLabel.Text = 'Ancho de banda de transición:';


% Create dBLabel
app.dBLabel = uilabel(app.SignalFilterPanel);
app.dBLabel.HorizontalAlignment = 'right';
app.dBLabel.Position = [179 110 83 22];
app.dBLabel.Text = 'dB';


% Create RipPasoEditField
app.RipPasoEditField = uieditfield(app.SignalFilterPanel, 'numeric');
app.RipPasoEditField.Position = [195 109 43 23];
app.RipPasoEditField.Value = 1;


% Create HzEditFieldLabel_3
app.HzEditFieldLabel_3 = uilabel(app.SignalFilterPanel);
app.HzEditFieldLabel_3.HorizontalAlignment = 'right';
app.HzEditFieldLabel_3.Position = [180 195 83 22];
app.HzEditFieldLabel_3.Text = 'Hz';


% Create FcorteEditField
app.FcorteEditField = uieditfield(app.SignalFilterPanel, 'numeric');
app.FcorteEditField.Position = [196 194 43 23];
app.FcorteEditField.Value = 4000;


% Create dBLabel_2
app.dBLabel_2 = uilabel(app.SignalFilterPanel);
app.dBLabel_2.HorizontalAlignment = 'right';
app.dBLabel_2.Position = [180 71 83 22];
app.dBLabel_2.Text = 'dB';


% Create AtCorteEditField
app.AtCorteEditField = uieditfield(app.SignalFilterPanel, 'numeric');
app.AtCorteEditField.Position = [196 70 43 23];
app.AtCorteEditField.Value = 60;


% Create RipPasoLabel
app.RipPasoLabel = uilabel(app.SignalFilterPanel);
app.RipPasoLabel.Position = [13 109 146 22];
app.RipPasoLabel.Text = 'Rizado en banda de paso:';
```

```
% Create AtCorteLabel
app.AtCorteLabel = uilabel(app.SignalFilterPanel);
app.AtCorteLabel.Position = [13 70 182 22];
app.AtCorteLabel.Text = 'Atenuación en la banda de corte:';


% Create NombredearchivoLabel
app.NombredearchivoLabel = uilabel(app.UIFigure);
app.NombredearchivoLabel.HorizontalAlignment = 'right';
app.NombredearchivoLabel.Position = [689 82 114 22];
app.NombredearchivoLabel.Text = 'Nombre de archivo: ';


            % Create FileNameEditField
            app.FileNameEditField = uieditfield(app.UIFigure, 'text');
            app.FileNameEditField.Position = [818 82 100 22];


            % Create GUARDARButton
            app.GUARDARButton = uibutton(app.UIFigure, 'push');
            app.GUARDARButton.ButtonPushedFcn = createCallbackFcn(app,
@GUARDARButtonPushed, true);
            app.GUARDARButton.Position = [766 42 100 22];
            app.GUARDARButton.Text = 'GUARDAR';


            % Create wavLabel
            app.wavLabel = uilabel(app.UIFigure);
            app.wavLabel.Position = [931 82 30 22];
            app.wavLabel.Text = '.wav';


            % Create GuardadoLabel
            app.GuardadoLabel = uilabel(app.UIFigure);
            app.GuardadoLabel.Position = [714 9 190 22];
            app.GuardadoLabel.Text = '';


            % Create PotenciaTextLabel
            app.PotenciaTextLabel = uilabel(app.UIFigure);
            app.PotenciaTextLabel.Position = [689 160 157 22];
            app.PotenciaTextLabel.Text = 'POTENCIA DE LA SEÑAL =';


            % Create PotenciaLabel
            app.PotenciaLabel = uilabel(app.UIFigure);
            app.PotenciaLabel.Position = [865 160 77 22];
            app.PotenciaLabel.Text = '-';
```

```
            % Show the figure after all components are created
            app.UIFigure.Visible = 'on';
        end
    end
```