ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)

Máster en Ingeniería en Tecnologías de Telecomunicación

Análisis y Evaluación de Sistemas de Autenticación para Smart Personal Assistants

Autor
Cayetano Valero Amores

Dirigido por
Dr. Gregorio López López
Dr. Guillermo Suárez-Tangil

Madrid
Julio 2021

Cayetano Valero Amores, declara bajo su responsabilidad, que el Proyecto con título **Análisis y Evaluación de Sistemas de Autenticación para Smart Personal Assistants** presentado en la ETS de Ingeniería (ICAI) de la Universidad Pontificia Comillas en el curso académico 2020/21 es de su 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.: ..... Fecha: 11 / 07 / 2021

Autoriza la entrega:

Los Directores del Proyecto

Dr. Gregorio López López
Dr. Guillermo Suárez-Tangil

Fdo.: ..... Fdo.: .....

Fecha: 11 / 07 / 2021

V. B. del Coordinador de Proyectos

Javier Matanza Domingo

Fdo.: ..... Fecha: ..... / ..... / .....

AUTORIZACIÓN PARA LA DIGITALIZACIÓN, DEPÓSITO Y DIVULGACIÓN EN RED DE PROYECTOS FIN DE GRADO, FIN DE MÁSTER, TESINAS O MEMORIAS DE BACHILLERATO

1º. Declaración de la autoría y acreditación de la misma.

El autor D. Cayetano Valero Amores **DECLARA** ser el titular de los derechos de propiedad intelectual de la obra: Análisis y Evaluación de Sistemas de Autenticación para *Smart Personal Assistants*, que ésta es una obra original, y que ostenta la condición de autor en el sentido que otorga la Ley de Propiedad Intelectual.

2º. Objeto y fines de la cesión.

Con el fin de dar la máxima difusión a la obra citada a través del Repositorio institucional de la Universidad, el autor **CEDE** a la Universidad Pontificia Comillas, los derechos de digitalización, de archivo, de reproducción, de distribución y de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, de comunicación pública, incluido el derecho de puesta a disposición electrónica, tal y como se describen en la Ley de Propiedad Intelectual. El derecho de transformación se cede a los únicos efectos de lo dispuesto en la letra a) del apartado siguiente.

3º. Condiciones de la cesión y acceso

Sin perjuicio de la titularidad de la obra, que sigue correspondiendo a su autor, la cesión de derechos contemplada en esta licencia habilita para:

(a) Transformarla con el fin de adaptarla a cualquier tecnología que permita incorporarla a internet y hacerla accesible; incorporar metadatos para realizar el registro de la obra e incorporar "marcas de agua" o cualquier otro sistema de seguridad o de protección.

(b) Reproducirla en un soporte digital para su incorporación a una base de datos electrónica, incluyendo el derecho de reproducir y almacenar la obra en servidores, a los efectos de garantizar su seguridad, conservación y preservar el formato.

(c) Comunicarla, por defecto, a través de un archivo institucional abierto, accesible de modo libre y gratuito a través de internet.

(d) Cualquier otra forma de acceso (restringido, embargado, cerrado) deberá solicitarse expresamente y obedecer a causas justificadas.

(e) Asignar por defecto a estos trabajos una licencia Creative Commons.

(f) Asignar por defecto a estos trabajos un HANDLE (URL *persistente*).

4º. Derechos del autor.

El autor, en tanto que titular de una obra tiene derecho a:

(a) Que la Universidad identifique claramente su nombre como autor de la misma

(b) Comunicar y dar publicidad a la obra en la versión que ceda y en otras posteriores a través de cualquier medio.

(c) Solicitar la retirada de la obra del repositorio por causa justificada.

(d) Recibir notificación fehaciente de cualquier reclamación que puedan formular terceras personas en relación con la obra y, en particular, de reclamaciones relativas a los derechos de propiedad intelectual sobre ella.

5º. Deberes del autor.

El autor se compromete a:

(a) Garantizar que el compromiso que adquiere mediante el presente escrito no infringe ningún derecho de terceros, ya sean de propiedad industrial, intelectual o cualquier otro.

(b) Garantizar que el contenido de las obras no atenta contra los derechos al honor, a la intimidad y a la imagen de terceros.

(c) Asumir toda reclamación o responsabilidad, incluyendo las indemnizaciones por daños, que pudieran ejercitarse contra la Universidad por terceros que vieran infringidos sus derechos e intereses a causa de la cesión.

(d) Asumir la responsabilidad en el caso de que las instituciones fueran condenadas por infracción de derechos derivada de las obras objeto de la cesión.

6º. Fines y funcionamiento del Repositorio Institucional.

La obra se pondrá a disposición de los usuarios para que hagan de ella un uso justo y respetuoso con los derechos del autor, según lo permitido por la legislación aplicable, y con fines de estudio, investigación, o cualquier otro fin lícito. Con dicha finalidad, la Universidad asume los siguientes deberes y se reserva las siguientes facultades:

- La Universidad informará a los usuarios del archivo sobre los usos permitidos, y no garantiza ni asume responsabilidad alguna por otras formas en que los usuarios hagan un uso posterior de las obras no conforme con la legislación vigente. El uso posterior, más allá de la copia privada, requerirá que se cite la fuente y se reconozca la autoría, que no se obtenga beneficio comercial, y que no se realicen obras derivadas.
- La Universidad no revisará el contenido de las obras, que en todo caso permanecerá bajo la responsabilidad exclusive del autor y no estará obligada a ejercitar acciones legales en nombre del autor en el supuesto de infracciones a derechos de propiedad intelectual derivados del depósito y archivo de las obras. El autor renuncia a cualquier reclamación frente a la Universidad por las formas no ajustadas a la legislación vigente en que los usuarios hagan uso de las obras.
- La Universidad adoptará las medidas necesarias para la preservación de la obra en un futuro.
- La Universidad se reserva la facultad de retirar la obra, previa notificación al autor, en supuestos suficientemente justificados, o en caso de reclamaciones de terceros.

Madrid, a 11 de Julio de 2021

ACEPTA

Fdo.: .....

Motivos para solicitar el acceso restringido, cerrado o embargado del trabajo en el Repositorio Institucional:

Análisis y Evaluación de Sistemas de Autenticación para Smart Personal Assistants

Máster en Ingeniería de Telecomunicación

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
Universidad Pontifica Comillas

Autor:
Cayetano Valero Amores

Directores:
Dr. Gregorio López López
Dr. Guillermo Suárez-Tangil

Julio, 2021

Resumen

Palabras clave

Asistentes personales, Seguridad, Privacidad, Clasificación de voz, Inteligencia Artificial, Python

Introducción

En los últimos años, gracias a la expansión de la Inteligencia Artificial, se han desarrollado y mejorado diferentes esquemas de interacción entre los humanos y los dispositivos electrónicos. Uno de los más populares en la actualidad es la interacción por voz, en la que los usuarios son capaces de comunicar las instrucciones que deseen a un dispositivo a través de comandos de voz, permitiendo una interacción más rápida y natural [1].

Una de las tecnologías que más ha contribuido a la mejora de los sistemas de interacción por voz gracias a su penetración en la vida de las personas [2] es la de los asistentes personales inteligentes o *Smart Personal Assistant* (SPA). Los SPA permiten a un usuario realizar gran cantidad de tareas utilizando su voz como medio de transmisión de la información, por ejemplo, los SPA ofrecen la posibilidad de consultar información como el tiempo o el estado del tráfico, escuchar música, realizar llamadas de voz y vídeo, hacer compras *online* o controlar otros dispositivos como luces y termostatos inteligentes [3], [4].

El presente Trabajo Fin de Máster se ha desarrollado en el marco del proyecto europeo empoweRing and educAtion YoUng pEopLe for the internet by plAying (RAYUELA), cuyo objetivo es identificar y comprender los factores humanos y tecnológicos relacionados con el cibercrimen, además de proporcionar educación a los menores en el uso de las tecnologías, para concienciar de los riesgos asociados a su uso y que sean capaces de identificar y prevenir comportamientos malintencionados en la red. Como parte del proyecto RAYUELA, el trabajo se incluye dentro del paquete de trabajo *Technology assessment and IT threat landscape*, cuyo objetivo es identificar los dispositivos conectados utilizados por menores y evaluar las amenazas y vulnerabilidades que les afectan.

Objetivos

En el contexto descrito, el objetivo general del Trabajo Fin de Máster es:

- Analizar y evaluar los riesgos y amenazas de seguridad y privacidad asociados a SPA, prestando especial atención a la autenticación por voz. Para conseguirlo, se han planteado cinco objetivos específicos:
    - Estudio de las tecnologías utilizadas en el ecosistema de los asistentes personales.
    - Análisis de los problemas de seguridad en el ecosistema de los asistentes personales.
    - Estudio de seguridad y privacidad de los principales asistentes personales inteligentes disponibles en el mercado.
    - Estudio de las herramientas de clasificación de voz basadas en Inteligencia Artificial.
    - Desarrollo de un modelo de clasificación de voz que permita distinguir voces de niños y adultos.

Desarrollo y resultados obtenidos

El estudio se ha segmentado en tres partes fundamentales.

Durante la primera fase del proyecto se ha realizado un estudio de las publicaciones científicas existentes acerca de la seguridad en el ecosistema de los SPA, que ha permitido identificar los principales problemas de seguridad que afectan a este tipo de dispositivos para determinar las hipótesis iniciales que han establecido las líneas de investigación que se han seguido en el proyecto.

En la segunda fase se ha diseñado y llevado a cabo un conjunto de pruebas de seguridad y privacidad en tres SPA comerciales, el *HomePod Mini* de *Apple*, el *Google Home Mini* de *Google* y el *Echo Show 5* de *Amazon*, para validar los problemas hallados durante el estudio de la documentación científica, así como llevar a cabo evaluaciones de las pruebas diseñadas durante el proyecto, centradas en evaluar los sistemas de autenticación y la protección de usuarios menores en este tipo de dispositivos. Los resultados de las pruebas en los asistentes comerciales han puesto de manifiesto que los sistemas de autenticación en los SPA analizados son deficientes, incluso en los dispositivos que cuentan con capacidades de reconocimiento de voz, una simple grabación de un usuario permite a un posible actor malicioso suplantar al usuario legítimo, realizando consultas al asistente en su nombre, pudiendo incluso acceder a información personal o realizar pagos [Figuras 1 y 2].

Figura 1: Grabación del mensaje de activación del asistente personal

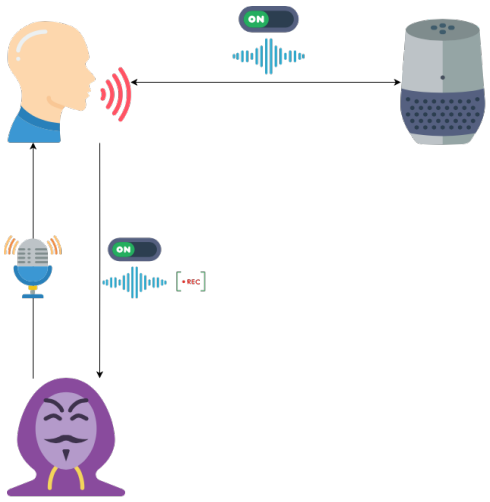
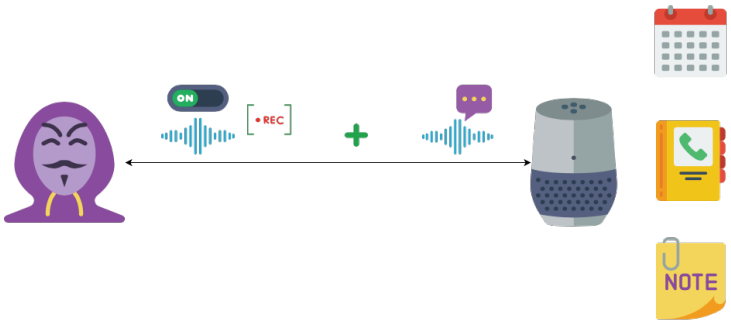


Figura 2: Ataque de suplantación con grabación del usuario legítimo



En cuanto a la protección de los usuarios menores en los dispositivos, se ha comprobado que ninguno de los tres SPA en las versiones analizadas cuentan con un sistema de perfiles de seguridad que permita definir un control de las funcionalidades adecuado a un uso seguro del dispositivo por parte de usuarios menores. Se han identificado dos líneas de trabajo complementarias que ayudarán a proteger a los usuarios menores en los SPA:

- Es necesario definir diferentes perfiles de seguridad preconfigurados con políticas de acceso que permitan controlar de forma sencilla pero precisa como los usuarios se relacionan con los dispositivos, tal como se hace en otras plataformas de contenido multimedia, como *Netflix* o *HBO*. En el ecosistema de los SPA la creación de perfiles de seguridad para menores es más compleja, ya que las funcionalidades de los dispositivos, y por tanto las necesidades de control, son más variadas y complejas que en una aplicación de reproducción multimedia, por lo que es necesario analizar exhaustivamente las características de los SPA para identificar y evaluar los riesgos que su uso por parte de un menor puede producir.

- Ante la necesidad de recurrir a las muestras de voz que el asistente captura en el momento de la invocación por parte de un usuario para autenticarle, es preciso desarrollar un sistema robusto que permita distinguir a usuarios menores y mayores de edad cada vez que se envía una consulta al asistente, para aplicar los perfiles de seguridad de forma automática. Con esto se favorecería el uso de las opciones de seguridad incluidas, ya que se aplicarían de forma transparente para los usuarios, tal como se aplican los mecanismos de autenticación en los *smartphones* actuales.

Los hallazgos realizados durante la segunda parte del proyecto han motivado el desarrollo llevado a cabo durante la tercera y última fase, en la que se han diseñado, optimizado y validado diversos modelos de clasificación de voz por grupos de edad basados en técnicas de Inteligencia Artificial utilizando conjuntos de datos abiertos en inglés y español, con el fin de comprobar el desempeño de este tipo de técnicas en el campo de la distinción automática de usuarios por edad basada en la voz. Con este estudio no se ha buscado desarrollar una herramienta infalible que pudiera ser usada en el ámbito comercial de forma directa, sino establecer un punto de partida en la línea de investigación de la clasificación de voz para que en futuros trabajos se continúen depurando los modelos, haciendo uso de nuevas técnicas de Inteligencia Artificial, conjuntos de datos alternativos y extrayendo características diferentes que permitan aumentar el rendimiento. Los resultados obtenidos en los sistemas de clasificación demuestran su utilidad en este campo, y abren posibles líneas de investigación para mejorar los sistemas en un futuro, para obtener herramientas robustas y fiables que puedan ser implementadas en los asistentes personales comerciales, que permitirán mejorar la seguridad y la usabilidad de los mismos.

Abstract

Keywords

Smart Personal Assistans, Security, Privacy, Voice classification, Artifical Intelligence, Python

Introduction

In recent years, as Artificial Intelligence has gained prominence, different interaction schemes between humans and electronic devices have been developed and improved. One of the most popular nowadays is voice interaction, in which users are able to communicate their instructions to a device through voice commands, allowing a faster and more natural interaction [1].

One of the technologies that has greatly contributed to the enhancement of voice interaction systems thanks to its penetration into people's lives is the Smart Personal Assistant (SPA). SPA enable a user to perform a large number of tasks using their voice as a means of conveying information, for example, SPA offer the ability to consult information such as weather or traffic status, listen to music, make voice and video calls, shop *online* or control other devices such as smart lights and intelligent thermostats [3], [4].

This Master's Thesis has been developed within framework of the European project *empoweRing and educAtion YoUng pEopLe for the internet by plAying* RAYUELA, whose objective is to identify and understand the human and technological factors related to cybercrime, as well as provide education to minors in the use of technology, in order to raise awareness about the risks associated with its use, making them able to detect and prevent malicious behaviours. As part of the RAUELA project, the work is included inside the work package *Technology assessment and IT threat landscape*, which aims to identify connected devices used by minors and assess the possible threats and vulnerabilities that affect them.

Objectives

In this context, the objective of the Master's Thesis is:

- Analyze and evaluate the security and privacy risks and threats associated with SPA, paying special attention to voice authentication. In order to achieve this goal, five specific objectives have been defined:
    - Research of the technologies used in the Smart Personal Assistant ecosystem.
    - Analysis of security issues in the personal assistant ecosystem.
    - Security and privacy study of the main Smart Personal Assistants available in the market.
    - Review of voice classification tools based on Artificial Intelligence.
    - Development of a voice classification model to differentiate between children and adult voices.

Development and results

The project has been divided into three fundamental parts.

During the first phase of the project, a research about existing scientific publications regarding security in the SPA ecosystem has been carried out. This research has made possible to identify the main security issues that affect the SPA environment to determine the initial hypotheses that have established the lines of research followed in the project.

In the second phase, a set of security and privacy tests have been designed and carried out on three commercial SPA, the *HomePod Mini* by *Apple*, the *Google Home Mini* by *Google* and the *Echo Show 5 by Amazon*, to validate the security problems found during the research of the scientific publications in the area, as well as to carry out evaluations of the tests designed during the project, focused on evaluating the authentication systems and the protection of minor users in this type of devices. The results of the tests have shown that the authentication systems in the SPA analyzed are deficient, even in devices that have voice recognition capabilities, a simple voice recording of a user allows a possible malicious actor to impersonate the legitimate user, making queries to the assistant on his behalf, even being able to access personal information or make payments [Figures 1 and 2].

Figure 1: Recording of the personal assistant wake-up message

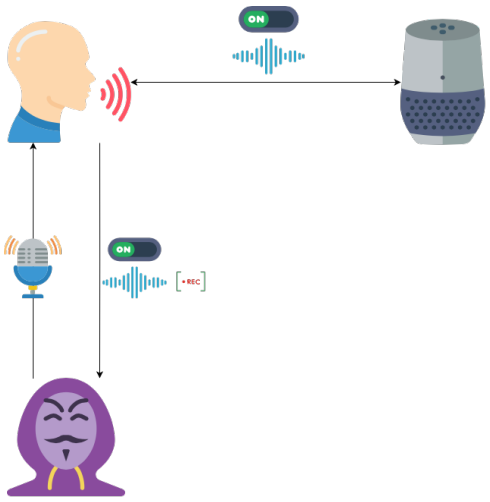
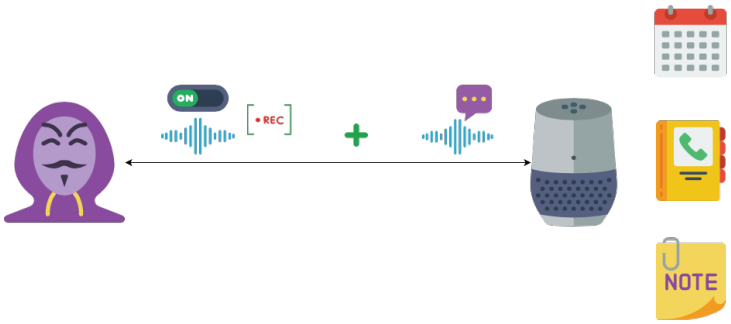


Figure 2: Impersonation attack with recording of a legitimate user



As for the protection of minors on these devices, it has been found that none of the three SPA in the versions analyzed have a security profile system that allows to define a proper functionalities control suitable for a secure use of the device by minor users. Two complementary lines of work have been identified to help protect minor users in SPA:

- It is necessary to define different preconfigured security profiles with access policies that allow simple but precise control of how users interact with the devices, as is done in other multimedia content platforms, such as Netflix or HBO. In the SPA ecosystem, the creation of security profiles for minors is more complex, since the functionalities of the devices, and therefore the control needs, are more varied and complex than in a multimedia playback application, so it is necessary to thoroughly analyze the characteristics of SPAs to identify and assess the risks that their use by a minor may produce.
- Given the need to rely on the voice samples that the assistant captures at the time of being invoked by a user to authenticate him, it is necessary to develop a robust system that allows to distinguish between minor and adult users each time a query is sent to the assistant, in order to apply security profiles automatically. This would facilitate the use of the security options included, since they would be applied transparently to users, just as the authentication mechanisms are applied in current smartphones.

The findings made during the second part of the project have motivated the development conducted during the third and last phase, in which several models of voice classification by age groups based on Artificial Intelligence techniques have been designed, optimized and validated using open data sets in English and Spanish, in order to test the performance of this type of techniques in the field of automatic voice-based age distinction of users. The aim of this research was not to develop an state-of-the art tool that could be used directly in the commercial environment, but to establish a baseline in the research of voice classification so that future work can continue to refine the models, making use of new Artificial Intelligence techniques, alternate datasets and extraction of different features to increase the performance. The results obtained in the classification systems demonstrate their usefulness in this field, and open possible lines of investigation to improve the systems in the future, to obtain robust and reliable tools that can be implemented in commercial personal assistants, which will improve their security and usability.

'I wish it need not have happened in my time,' said Frodo.
'So do I,' said Gandalf, 'and so do all who live to see such times.
But that is not for them to decide.
All we have to decide is what to do with the time that is given us'
J. R. R. Tolkien, The Fellowship of the Ring

Agradecimientos

Gracias a Gregorio y Guillermo por los consejos y el apoyo prestado durante el proyecto. Gracias por vuestro tiempo y esfuerzo en hacer de este proyecto lo que es hoy. Gracias por darme la oportunidad de colaborar en el proyecto RAYUELA, pudiendo aportar el saber hacer de la Universidad colaborando con grandes compañeros de viaje, implicados y trabajadores.

Gracias a ICAI y a sus profesores por estos últimos seis años, por la formación técnica y personal recibida, por fomentar el trabajo duro y el esfuerzo durante todos estos años, de los que me llevo los mejores compañeros que podría desear y que espero me acompañen durante muchos años más.

Gracias a mi familia, a mis padres y a mi hermano por acompañarme durante este viaje y por haberme ayudado todo lo posible para tener éxito en esta etapa.

Cayetano Valero Amores
Madrid, Julio de 2021

Índice general

Índice de figuras

Índice de tablas

Índice de fragmentos de código

Acrónimos

openSMILE *open-source Speech and Music Interpretation by Large-space Extraction*. 19, 69, 75–77

BIC *Bayesian Information Criterion*. 94

CMS *Cepstral Mean Substraction*. 10

EM *Expectation Maximization*. 11, 12

GMM *Gaussian Mixture Models*. 11, 12, 78, 93, 94, 96

ODS Objetivos de Desarrollo Sostenible. 113

PDF *Probability Density Function*. 11

RAYUELA *empoweRing and educAtion YoUng pEopLe for the internet by plAying*. 2, 109

RBF *Radial Basis Function*. 13, 81

RF *Random Forest*. 11, 14, 78, 86, 87, 110

SPA *Smart Personal Assistants*. 1–8, 21, 23, 25, 31, 42, 56, 101, 109, 111, 113

SVM *Support Vector Machines*. 11, 13, 14, 78–81, 86, 88, 110

TTS *Text To Speech*. 22

Capítulo 1

Introducción

En los últimos años, gracias a la expansión de la Inteligencia Artificial, se han desarrollado y mejorado diferentes esquemas de interacción entre los humanos y los dispositivos electrónicos. Uno de los más populares en la actualidad es la interacción por voz, en la que los usuarios son capaces de comunicar las instrucciones que deseen a un dispositivo a través de comandos de voz, permitiendo una interacción más rápida y natural [1].

Una de las tecnologías que más ha contribuido a la mejora de los sistemas de interacción por voz gracias a su penetración en la vida de las personas [2] es la de los asistentes personales inteligentes o *Smart Personal Assistants* (SPA). Los SPA permiten a un usuario realizar gran cantidad de tareas utilizando su voz como medio de transmisión de la información, por ejemplo, los SPA ofrecen la posibilidad de consultar información como el tiempo o el estado del tráfico, escuchar música, realizar llamadas de voz y vídeo, hacer compras *online* o controlar otros dispositivos como luces y termostatos inteligentes [3], [4].

Los SPA se diferencian de los sistemas de interacción por voz tradicionales en el procesamiento que se realiza con los comandos recogidos del usuario. Mientras que en un sistema tradicional el usuario tiene que invocar un comando exacto con una pronunciación adecuada para activar una función, los SPA se valen de los avances en el procesado de lenguaje natural para "entender" los comandos de voz del usuario, extrayendo la intención de sus mensajes para dar la mejor respuesta a su pregunta, sin necesidad de que estas sigan una estructura rígida o se pronuncien de una forma determinada.

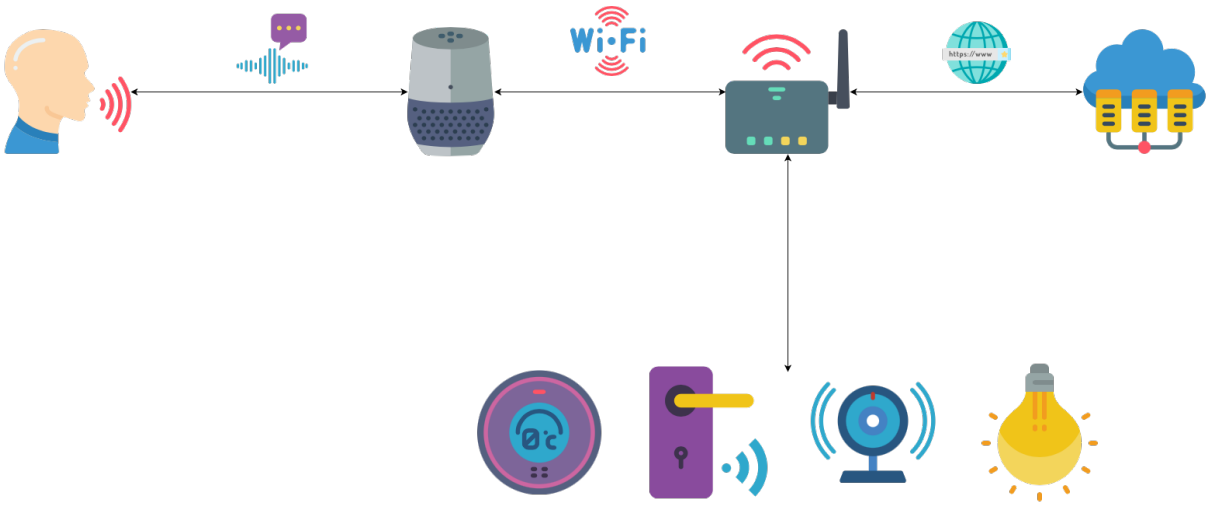
Para hacer posible las funcionalidades descritas anteriormente, el ecosistema de los SPA consta de varios elementos, indicados en un esquema simplificado en la Figura 1.1:

- Dispositivo con agente SPA. Varias familias de dispositivos electrónicos tienen compatibilidad con asistentes personales, como altavoces, *smartphones*, *PCs* y *wearables*. Sus funciones son recoger los comandos del usuario para enviarlos a la nube del proveedor y reproducir las respuestas que llegan de la nube.

- Nube del proveedor. En este componente reside toda la inteligencia del asistente. Es el lugar donde se realiza el procesado de las peticiones para extraer la intención del usuario y buscar la respuesta que mejor se adapte a su pregunta.

- Nubes de terceros y nubes de accesorios. Los usuarios pueden interactuar con elementos externos al proveedor del SPA, por ejemplo, con *skills* de terceros o dispositivos inteligentes de otros fabricantes. Para alcanzar a estos elementos externos, el SPA está conectado con las nubes de los proveedores de dichos elementos.

Figura 1.1: Arquitectura del ecosistema SPA



[1]

El crecimiento de las tecnologías en el ámbito del hogar conectado [5], [6], [7] hace que cada vez más dispositivos tradicionales que se encuentran en el hogar, como bombillas o termostatos, cuenten con una versión "inteligente" con capacidad de conexión a Internet [8], [9], que permite el control remoto del dispositivo, así como la automatización de ciertas tareas y la programación de actividades en base a los parámetros que el usuario configure.

Por su facilidad de uso y sus capacidades de integración, los SPA se han posicionado como los elementos centrales del hogar conectado [Figura 1.1], actuando como punto de conexión entre el usuario y el resto de los dispositivos. Cada vez más fabricantes de dispositivos inteligentes integran sus productos en los ecosistemas de los SPA [10], [11], [12] para ofrecer una experiencia de uso más transparente y un proceso de configuración más sencillo a sus clientes.

Al ser el punto que unifica la comunicación con los dispositivos inteligentes, el SPA se convierte en un objetivo de alto valor para un posible actor malicioso, ya que el acceso al asistente permitiría interactuar con los dispositivos y la información asociados al SPA.

La información que se recoge y procesa en el ecosistema de los SPA es uno de los activos más importantes desde el punto de vista de los usuarios, debido a la naturaleza de los datos que se registran y envían. Entre los datos personales que un SPA procesa se incluyen datos de la cuenta del usuario, datos de interacción con los servicios y las propias grabaciones de voz del usuario [13], [14], [15].

La obtención y el procesado de la información de los usuarios está relacionada estrechamente con la funcionalidad que los SPA ofrecen, así como con la configuración que el usuario pueda realizar para ajustar la recogida de datos, por lo que es importante conocer las características de los dispositivos que los mayores proveedores ofrecen con el fin de realizar un análisis de seguridad adecuado. A escala global, y exceptuando los asistentes personales sólo disponibles en el mercado chino, los asistentes personales más usados son *Assistant* de *Google*, *Alexa* de *Amazon* y *Siri* de *Apple* [16], [17].

El estudio se enmarca dentro del proyecto europeo *empoweRing and educAtion YoUng pEopLe for the internet by plAying* (RAYUELA) [18] [Figura 1.2], cuyo objetivo es analizar y comprender los factores humanos y tecnológicos relacionados con el cibercrimen, a la vez que se educa a los menores para que tomen conciencia de los riesgos y amenazas existentes en *Internet* para que puedan identificar y prevenir comportamientos peligrosos en la red. Dentro del proyecto se desarrollarán diferentes paquetes de trabajo centrados en aspectos clave del proyecto (pueden consultarse los paquetes de trabajo en la sección *Work Packages* en [18]). El estudio se enmarca concretamente dentro del paquete de trabajo número dos, *Technology assessment and IT threat landscape*, cuyo objetivo es analizar las amenazas y vulnerabilidades presentes en los dispositivos tecnológicos actuales y estudiar su relación con los menores. Los avances realizados en el paquete de trabajo en los que se ha contribuido con este proyecto han dado lugar a una publicación en la que se analiza la seguridad y privacidad de los principales dispositivos IoT utilizados por jóvenes [19], que ha sido publicada en la sexta edición de las Jornadas Nacionales de Investigación en Ciberseguridad, celebradas en Junio de 2021 y organizadas por la Universidad de Castilla-La Mancha.

[1] Iconos creados por *Smash icons* en flaticon.com

Figura 1.2: Logo del proyecto RAYUELA

1.1. Motivación

En un contexto tan cambiante y complejo, donde surgen nuevas técnicas y tecnologías a gran velocidad y la integración de los asistentes personales con otros dispositivos electrónicos en ecosistemas comunes es cada vez mayor, resulta esencial entender y analizar el contexto en el que se trabaja, para tener clara la composición del ecosistema, identificando que herramientas y datos entran en juego, ya que sólo así se podrán determinar y valorar correctamente los posibles riesgos de seguridad a los que se exponen los usuarios para tratar de minimizarlos.

Los sistemas de autenticación de voz representan la puerta de entrada, tanto a usuarios legítimos como a posibles actores maliciosos, a todo el ecosistema de dispositivos conectados a los asistentes personales. Como se ha visto en el análisis realizado del ecosistema de los SPA, la autenticación y autorización representan un punto débil en cuanto a la seguridad en la arquitectura del ecosistema de asistentes personales, por lo que con este proyecto se busca estudiar y analizar la seguridad y privacidad en modelos comerciales de asistentes personales para corroborar la situación expuesta, así como desarrollar y validar técnicas de clasificación de voz basadas en Inteligencia Artificial para permitir que las personas puedan hacer un uso más seguro de los SPA, sin comprometer su usabilidad y flexibilidad, en base a los hallazgos que se produzcan en el estudio de los asistentes comerciales.

1.2. Objetivos

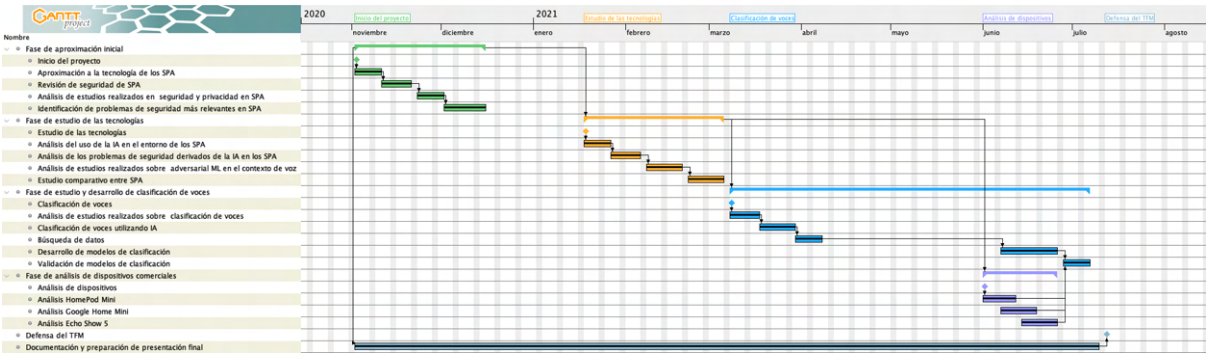
El objetivo general del proyecto es:

- Analizar y evaluar los riesgos y amenazas de seguridad y privacidad asociados a SPA, prestando especial atención a la autenticación por voz. Para conseguirlo, se han planteado cinco objetivos específicos:
    - Estudio de las tecnologías utilizadas en el ecosistema de los asistentes personales.
    - Análisis de los problemas de seguridad en el ecosistema de los asistentes personales.
    - Estudio de seguridad y privacidad de los principales asistentes personales inteligentes disponibles en el mercado.
    - Estudio de las herramientas de clasificación de voz basadas en Inteligencia Artificial.
    - Desarrollo de un modelo de clasificación de voz que permita distinguir voces de niños y adultos.

1.3. Planificación

Para la realización del proyecto se ha seguido un enfoque de estudio clásico, mostrado en la Figura 1.3, cuya primera parte ha estado centrada en el estudio y comprensión de las tecnologías del entorno en el que se desarrolla el proyecto. Durante la segunda parte del proyecto se han estudiado las tecnologías específicas del ecosistema de los asistentes personales, para proceder durante la tercera y cuarta fase a realizar el análisis de los asistentes personales comerciales y el desarrollo del clasificador de voz por grupos de edad basado en Inteligencia Artificial.

Figura 1.3: Planificación del Trabajo Fin de Máster



1.4. Organización de la memoria

La memoria del proyecto está organizada en seis capítulos.

En el Capítulo 1 se recoge la introducción al ecosistema de los asistente personales inteligentes, así como la motivación, objetivos y planificación del proyecto.

En el Capítulo 2 se trata el estado del arte, en el que se detallan los estudios acerca de la seguridad en SPA que se han investigado, resaltando los principales problemas de seguridad que se han identificado en el ecosistema de los SPA. También se describen en este capítulo las herramientas, librerías y conjuntos de datos que se han usado en el desarrollo del proyecto.

En el Capítulo 3 se incluye el desarrollo del estudio de seguridad y privacidad realizado sobre asistente personales comerciales para identificar posibles problemas de seguridad que den sentido al desarrollo posterior del clasificador.

En el Capítulo 4 se cubre el desarrollo del clasificador de voces de niños y adultos basado en técnicas de Inteligencia Artificial, cubriendo todas las fases del desarrollo utilizando el código como apoyo a la explicación escrita. En este capítulo se encuentran detalladas todas las partes del proceso, desde el preprocesamiento de los datos hasta la creación, optimización y validación de los modelos.

En el Capítulo 5 se analizan los resultados obtenidos en el estudio de los asistentes personales comerciales analizados y en el desarrollo del clasificador de voces de niños y adultos.

Finalmente, en el Capítulo 6 se exponen las conclusiones del trabajo realizado y se presentan posibles líneas de investigación futuras que han surgido a lo largo del desarrollo del proyecto.

Capítulo 2

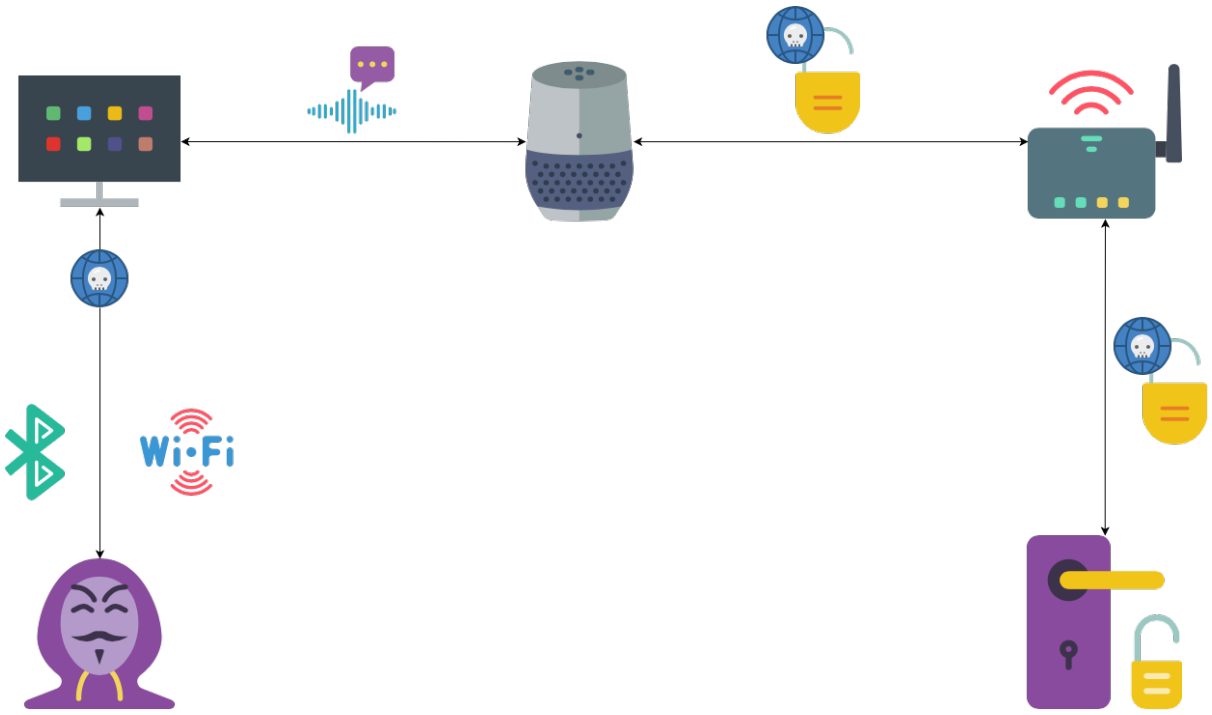
Estado del arte

2.1. Aspectos generales sobre seguridad y privacidad en SPA

Los asistentes personales tienen una serie de problemas de seguridad derivados de su arquitectura y su funcionamiento. Estos problemas se han estudiado en diferentes análisis, para comprobar como pueden ser utilizados por un posible actor malicioso y que impacto podrían causar en los usuarios. Principalmente, los problemas de seguridad en los SPA son debidos a la naturaleza abierta del canal de comunicación entre el usuario y asistente, las tecnologías y dispositivos subyacentes que se integran en la arquitectura y el uso de la Inteligencia Artificial, en la que se basan algunas de las funcionalidades de los SPA. Los riesgos descritos anteriormente dan lugar a diversos problemas de seguridad que se pueden clasificar en cuatro categorías [13]:

Autenticación débil Los SPA basan su funcionalidad en un modo escucha permanente. El asistente está escuchando en todo momento, esperando a que el usuario lo active e invoque una función concreta. La activación se realiza mediante una palabra clave que el usuario transmite al SPA y que hará que entre en el modo de funcionamiento activo, durante el cual el asistente graba la voz del usuario para procesarla y generar una respuesta. Por tanto, el único método de autenticación disponible en un SPA es la palabra el usuario utiliza para activarlo, que es escogida de entre un conjunto predefinido y limitado, por lo que se facilita a un atacante la posibilidad de adivinarla y poder interactuar con el asistente como un usuario legítimo. Al estar el asistente siempre encendido, se aumenta la superficie temporal en la que un actor malicioso podría atacar al SPA, pudiendo incluso utilizar dispositivos electrónicos cercanos al equipo en el que el SPA está instalado para inyectar comandos sin estar presente físicamente y realizar acciones como desbloquear las puertas de la casa [Figura 2.1] o realizar compras en nombre del usuario [20].

Figura 2.1: Ejemplo de un ataque remoto a un SPA



[1]

Autorización débil Los asistentes personales cuentan con funciones de distinción de usuarios en entornos familiares, para adaptar su respuesta en función del usuario que invoque la acción. Algunos de los SPA más usados [16], como *Alexa* o *Google Assistant* incluyen esta funcionalidad, pero no especifican que la distinción sea fiable, y alertan del hecho de que un usuario con una voz similar a otro pueda acceder a su información personal, haciéndose pasar por él [21].

Adversarial AI El uso de la Inteligencia Artificial en los SPA conlleva una serie de riesgos propios de esta tecnología, como el uso de la propia Inteligencia Artificial para atacar a los modelos de reconocimiento de voz [22]. Se presupone que los entornos en los que se usarán los modelos son seguros, pero debido a la naturaleza abierta del canal de comunicación entre el usuario y el SPA, existe la posibilidad de que un actor malicioso pueda inyectar muestras especialmente diseñadas para provocar un comportamiento potencialmente no deseado en el SPA, pudiendo forzar acciones maliciosas, como descargar *malware* en el dispositivo en el que el SPA está instalado accediendo a una URL, o provocar una denegación de servicio [23].

Uso de skills de terceros Las *skills* de terceros extienden la funcionalidad de un SPA, permitiendo que los usuarios puedan incluir en sus asistentes desarrollos de terceros para que puedan ser invocados de igual forma que se haría con una función nativa del asistente. Sin embargo, esto aumenta el riesgo al que se expone un usuario, ya que la superficie de ataque frente a un actor malicioso aumenta, pudiendo un atacante crear *skills* maliciosas con una pronunciación similar a una legítima para secuestrar las peticiones del usuario dirigidas a la *skill* legítima, o pudiendo crear *skills* que tomen el control de servicios que normalmente están reservados al SPA, como el inicio y la terminación de las interacciones con las *skills*, para hacer creer al usuario que una *skill* se ha ejecutado correctamente y el SPA ha dejado de escuchar, mientras que en realidad la *skill* maliciosa sigue activa, grabando y enviando información al atacante sin que el usuario sea consciente [24].

[1] Iconos creados por *Smash icons* en flaticon.com

En las tres primeras categorías anteriormente descritas, la interacción por voz del usuario con el SPA representa el punto débil en la arquitectura, ya que los mecanismos de identificación, autenticación y autorización no son robustos, y el SPA permanece a la escucha todo el tiempo, haciendo más factible un posible ataque. El mecanismo de defensa principal contra estos problemas de seguridad es la autenticación basada en la propia voz de los usuarios, que es el dato biométrico que todos los SPA recogen y procesan, independientemente de la plataforma o el dispositivo sobre el que se ejecuten.

Además, el escenario de los SPA presenta retos adicionales relacionados con el uso de muestras adversarias contra los modelos de reconocimiento de voz. Dichos modelos están diseñados teniendo en cuenta que todas las muestras que llegan para realizar inferencias son benignas y no están alteradas de forma artificial para provocar un comportamiento no deseado del modelo, pero estudios recientes han demostrado que se pueden modificar las muestras que procesan los modelos, alterando la interpretación que el modelo hace de los comandos de voz [22] o realizando una denegación de servicio, haciendo que el asistente no pueda recibir los comandos de un usuario [23], tal como se ha expuesto anteriormente.

Este escenario tan complejo requiere de mecanismos de defensa basados en inteligencia artificial, que ayudará a la hora de clasificar las muestras de voz que llegan al modelo antes de que sean procesadas, para diferenciar entre muestras sintéticas y humanas, con el fin de mitigar los ataques de *adversarial AI*.

Con respecto al uso de Inteligencia Artificial para llevar a cabo tareas de clasificación en el contexto de la voz se han realizado diferentes estudios que han tratado de diferenciar entre muestras de voz sintéticas y naturales, basándose en la aplicación de transformaciones en las muestras de voz y la posterior extracción de características, para obtener un conjunto de datos que permita a un modelo distinguir de forma precisa entre los dos tipos de muestras. De especial relevancia son también los estudios realizados en el ámbito de la clasificación por edades de los usuarios a través de muestras de voz, ya que sientan las bases del desarrollo que se llevará a cabo en el proyecto.

2.2. Sistemas de autenticación en SPA

Existen diferentes posibilidades a la hora de autenticar a un usuario cuando interactúa con un SPA, que pueden ir desde un nivel básico hasta otros esquemas que combinan información del usuario para ofrecer una autenticación más segura. Si bien estos esquemas de autenticación combinados son más robustos que el básico, esto no los convierte en seguros.

La autenticación de usuarios en los SPA se basa en la muestra de voz recogida cuando el usuario interactúa con el asistente haciendo uso de la palabra de activación correspondiente. La palabra de activación (*wake-up word*) sirve para que el asistente pase del modo escucha, que está activo de forma permanente a no ser que el usuario apague los micrófonos del SPA, al modo activo, en el que procesará las peticiones del usuario.

El esquema de autenticación más básico disponible en los SPA es la personalización de la palabra de activación que servirá para encender al asistente. El usuario puede escoger la palabra de activación de entre las presentes en un grupo predefinido. Las palabras de activación varían en función del fabricante, pero el conjunto de palabras que se pueden seleccionar es muy limitado [25], lo que aumenta la posibilidad de un atacante de adivinar la palabra de activación, que le permitirá interactuar con el asistente de la misma forma que lo haría un usuario legítimo.

Los asistentes cuentan con un esquema de autenticación más avanzado, que añade una capa de seguridad adicional a la palabra de activación. Mediante este esquema, se realiza una identificación del usuario a través del reconocimiento de su voz, comparando la muestra de voz que se recoge en el momento de la activación del asistente con el modelo que se ha creado para el usuario. Esta capacidad está presente en los SPA más extendidos, pero se ofrece como una funcionalidad que ayuda en la particularización de las respuestas para un usuario de forma transparente en función de quien haya invocado al asistente con la palabra de activación. No se oferta como una función de seguridad debido a las tasas de falsos positivos que se producen en usuarios con voces similares, remarcando este hecho incluso en la documentación del fabricante [21], tal como se expuso en el apartado de *Autorización débil* en la sección 2.1. Dependiendo de la configuración que un usuario haya realizado en el asistente en cuanto al acceso a datos personales y otras funcionalidades como la gestión de pagos, la confusión por parte del SPA al identificar a un usuario supondría poder acceder a su información personal, pudiendo enviar correos electrónicos o visualizar los eventos de su calendario, o incluso realizar pagos en su nombre.

También es posible un modo de autenticación alternativo que añade una nueva dimensión al método descrito anteriormente, el reconocimiento facial del usuario (*Face Match* en *Google Assistant*). En los dispositivos compatibles que cuentan con cámaras (*Nest Hub Max* [26] para el caso de *Google Assistant*), un usuario puede tomar diferentes fotografías con un dispositivo móvil para generar un modelo que se compartirá con el

SPA. Cuando el usuario se encuentra en el rango de visión de la cámara y realice una petición al asistente invocándolo con la palabra de activación, la cámara del dispositivo también se encenderá para realizar un reconocimiento facial del usuario que comenzó la petición, comprobando si se corresponde con el usuario al que se le asocia la muestra de voz de la petición. De nuevo, esta opción se oferta como una característica adicional para personalizar las respuestas que el asistente proporciona en función del usuario que realice las preguntas. En la documentación del fabricante, se destacan especialmente situaciones en las que el asistente puede ser engañado con una fotografía de un usuario registrado con *Face Match* [27], por lo que el nivel de seguridad que proporciona la solución en un entorno donde el acceso físico al SPA no esté estrictamente controlado, pudiendo ser susceptible a interacciones por parte de usuarios maliciosos, es muy limitado. Un acceso de un actor malicioso que suplante la identidad de un usuario legítimo le permitiría recabar información personal, como mensajes de vídeo, recordatorios o eventos de su calendario.

2.3. Clasificación de voz utilizando Inteligencia Artificial

La Inteligencia Artifical ha abierto nuevos caminos a la hora de afrontar problemas complejos de optimización, predicción o clasificación, entre otros. Durante los últimos años se han producido avances remarcables en la investigación [28] que han mejorado en gran medida los resultados de las técnicas de Inteligencia Artificial, hasta el punto de convertirlas en la solución *de facto* para determinados escenarios o abriendo soluciones a problemas que, debido a su complejidad, no podían ser resueltos de forma clásica, aplicando un pensamiento basado en reglas lógicas. Los modelos matemáticos basados en Inteligencia Artificial proporcionan mecanismos para identificar características significativas de los datos de entrada de un problema, con el fin de realizar una generalización sobre los datos que permita identificar patrones o comportamientos en los mismos de una forma más eficiente.

En el campo de la seguridad en los *Smart Personal Assistants*, la Inteligencia Artificial proporciona nuevas aproximaciones para realizar tareas de autenticación y autorización de usuarios más robustas y eficientes. Como se vio anteriormente, los SPA están pensados para ser el centro del hogar conectado, siendo dispositivos apropiados para un entorno familiar, donde pueden necesitarse políticas de diferenciación a la hora de definir los permisos con los que cuentan determinados usuarios para interactuar con el SPA. La clasificación de usuarios en base a su voz utilizando IA permite aplicar estos controles de acceso de forma automática, resultando en una capa de seguridad que trabaja de forma transparente en función de la persona que esté haciendo uso del asistente. Aplicando estas técnicas se puede dar solución a escenarios como la restricción de la funcionalidad de pago automático desde el asistente para niños, o la desactivación del control de determinados dispositivos, como puertas o ventanas conectadas, para usuarios invitados o niños.

La clasificación de usuarios en base a su voz se lleva a cabo en dos dimensiones: género y edad. La clasificación por género puede ser considerada como un problema de clasificación con dos categorías de salida, masculina o femenina, mientras que la clasificación por edades puede verse como un problema de clasificación con múltiples intervalos discretos de salida, o como un problema de predicción donde el resultado es un valor de edad numérico continuo [29].

Para el caso particular del ecosistema SPA, es de especial relevancia la distinción por edad entre usuarios adultos y niños. Esta características representa la separación fundamental para determinar que funcionalidades entrañan un riesgo al ser usadas por niños (como pagos o interacción con dispositivos sensibles como cerraduras o vehículos conectados) y deberían estar restringidas.

La clasificación de usuarios por edad entraña una serie de dificultades. Es necesario un conjunto de datos amplio y etiquetado por clases. La voz de cada persona tiene parámetros que varían en función de características físicas como la edad o la altura, y otras características que no están necesariamente relacionadas con la edad. Estas variaciones tienen que estar recogidas en el conjunto de datos para poder realizar una generalización adecuada. Además, el conjunto de datos tiene que ser representativo, es decir, el número de ocurrencias de cada una de las clases tiene que estar balanceado para recoger correctamente las características de cada clase y no introducir sesgo en la clasificación posterior.

2.3.1. Características de la voz

Para realizar una correcta clasificación de usuarios en base a muestras de voz, es necesario extraer una serie de parámetros que permitan describir la muestra de voz de forma numérica.

Las características de la voz de niños y adultos son diferentes. Por ejemplo, el *pitch* de la voz de los niños se encuentra en una parte del espectro superior a la de los adultos, y los armónicos del *pitch* se encuentran más separados en frecuencia. También se ha observado que la duración de los fonemas es menor en niños que

en adultos [30]. Los métodos de análisis de la voz deberán ser adaptados para trabajar correctamente con ambos grupos de usuarios.

Idealmente, las características que se extraigan de las muestras de voz deben permitir identificar a un usuario sin verse afectadas por factores externos como niveles de ruido ambiental no elevados, el estado de salud del usuario, etc. Estas características deben ocurrir de forma regular durante el habla y deben ser difíciles de imitar.

Los trabajos de investigación sobre clasificación de usuarios en función de la voz se han centrado en los parámetros acústicos de la misma. Como punto de partida general, se pueden distinguir tres categorías principales de características acústicas que se pueden extraer de la voz, indicadas a continuación.

2.3.1.1. Características prosódicas

Son aquellas que describen la generación de la propia voz. Las más utilizadas son:

- Duración. Mide la duración de los sonidos.
- *Pitch*. Propiedad que permite ordenar voces en una escala relacionada con la frecuencia, identificando sonidos "altos" o "bajos".
- Energía. Representa la energía de la señal de voz, que viene dada por la siguiente expresión:

$$E = \int_0^t |x(t)|^2 \, dt$$

donde:

$x(t)$ es la señal de audio.

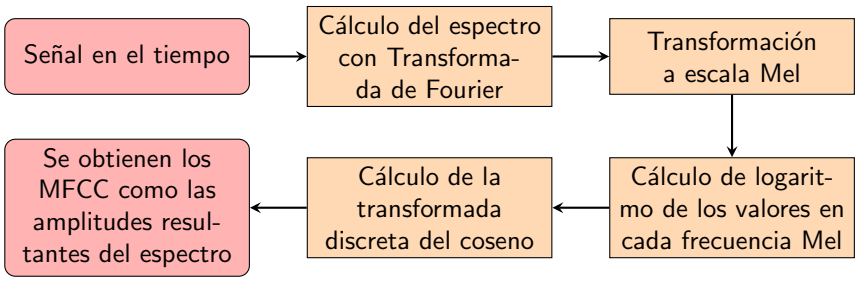
t es la duración de la señal.

2.3.1.2. Características espectrales

Características espectrales lineales Hacen referencias a los parámetros extraídos del espectro de frecuencia de la señal, calculado con la transformada de Fourier de la señal en el dominio del tiempo, representada sobre una escala lineal.

Características cepstrales Obtenidas de la transformación de Fourier del logaritmo del espectro de la señal. Son unas de las características más usadas en el campo del reconocimiento automático de usuarios. Existen características derivadas que se obtienen de la representación *ceptral* de la señal, siendo muy extendidos en el ámbito del reconocimiento de voz los coeficientes MFCC (*Mel-Frequency Cepstral Coefficients*), que son obtenidos realizando la transformada del coseno del logaritmo del espectro de la señal representado en una escala de frecuencia de Mel. Con los MFCC se obtiene una representación numérica compacta de la señal [Figura 2.2].

Figura 2.2: Cálculo de los coeficientes MFCC



Las variaciones en el canal afectan al rendimiento de las técnicas de reconocimiento de voz, por lo que es necesario normalizar las características extraídas en un intervalo de tiempo para que el rendimiento no se degrade cuando se trabaje en condiciones reales. Uno de los métodos más utilizados es el *Cepstral Mean Substraction* (CMS) [31].

Cepstral Mean Susbtraction CMS es una de las primeras técnicas que se propusieron para compensar los efectos que introduce la variabilidad de un canal sobre las características *cepstrales* de la señal de voz. Los vectores de características $\vec{c}_{cms}, i = 1, ..., N$ se extraen de los N vectores *cepstrales* de la señal de voz $y(n)$ restando la media *cepstral* a cada uno de los vectores *cepstrales* originales $\vec{c}_{y,i},\ i = 1, ..., N$ [32]:

$$\vec{c}_{cms,i} = \vec{c}_{y,i} - \vec{c}_{y,avg},\ i = 1, ..., N \tag{2.1}$$

donde:

$$\vec{c}_{y,avg} = \frac{1}{N}\sum_{i=1}^{N} \vec{c}_{y,i}$$

CMS se basa en el comportamiento del *cepstrum* ante distorsiones convolucionales, y la asunción de que el canal no cambia significativamente durante una muestra de voz. Una distorsión convolucional en el dominio del tiempo, como las que puede introducir un canal, se convierte en una distorsión aditiva en el dominio *cepstral*. Si denominamos $s(n)$ a la señal original y $h(n)$ al canal, la señal original es modificada tal como se indica a continuación:

$$y(n) = h(n) * s(n) \iff \vec{c}_y = \vec{c}_h + \vec{c}_s \tag{2.2}$$

donde $\vec{c}_y$, $\vec{c}_h$ y $\vec{c}_s$ representan las características *cepstrales* de la señal afectada por el canal, del propio canal y de la señal original respectivamente.

Si se calculan las medias *cepstrales* de cada uno de los componentes de la relación, se obtiene

$$\begin{aligned} \vec{c}_{y,avg} &\triangleq \frac{1}{N}\sum_{i=1}^{N} \vec{c}_{y,i} \\ &= \frac{1}{N}\sum_{i=1}^{N} \vec{c}_{h,i} + \frac{1}{N}\sum_{i=1}^{N} \vec{c}_{s,i} \\ &\triangleq \vec{c}_{h,avg} + \vec{c}_{s,avg} \end{aligned} \tag{2.3}$$

Asumiendo que el canal no varía durante la muestra de voz, el término $\frac{1}{N}\sum_{i=1}^{N} \vec{c}_{h,i}$ se convierte en $\vec{c}_h$, es decir, los componentes *cepstrales* del canal.

Si la distribución y variaciones de sonidos en la muestra de voz original $s(n)$ son tales que el espectro de la señal sea relativamente plano, los vectores de características *cepstrales* correspondientes tenderán a cero,

$$\vec{c}_{s,avg} \implies \vec{0} \tag{2.4}$$

Por tanto, la media *cepstral* de la señal afectada por el canal $y(n)$, tal como se ha definido en la ecuación 2.3 será:

$$\begin{aligned} \vec{c}_{y,avg} &= \vec{c}_{h,avg} + \vec{c}_{s,avg} \\ &= \vec{c}_h + \vec{0} \\ &= \vec{c}_h \end{aligned} \tag{2.5}$$

es decir, la media *cepstral* de la señal de salida será el *cepstrum* del canal. Para eliminar la media *cepstral* de cada uno de los vectores *cepstrales*, como se indica en la ecuación 2.1, bastará con restar el cepstrum del canal, lo que permite reescribir la ecuación 2.1 como:

$$\begin{aligned} \vec{c}_{cms,i} &= \vec{c}_{y,i} - \vec{c}_{y,avg} \\ &= (\vec{c}_h + \vec{c}_{s,i}) - \vec{c}_h \\ &= \vec{c}_{s,i} \end{aligned} \tag{2.6}$$

Los vectores *cepstrales* resultantes no dependen del canal, siendo invariantes al efecto de cualquier canal presente.

2.3.1.3. Características relacionadas con la calidad del sonido

Definen la claridad o pureza del sonido, como de fácil es distinguir sonidos diferentes en una grabación.

2.3.2. Algoritmos de clasificación

Existen diferentes tipos de algoritmos que han demostrado una gran aplicabilidad en el campo del procesamiento del audio, como son *Support Vector Machines* (SVM), *Random Forest* o *Gaussian Mixture Models* (GMM).

2.3.2.1. Clasificadores estadísticos (GMM)

Generalmente, las muestras de voz que se van a utilizar para desarrollar los modelos de clasificación contarán con duraciones diferentes. Para poder hacer uso de las muestras en el entrenamiento y validación de esta familia de modelos, es necesario que cuente con una longitud fija. Por tanto, las muestras deben representarse como vectores con un tamaño fijo. Para realizar esta adaptación, se utilizan funciones de densidad de probabilidad o *Probability Density Function* (PDF), que se aplican a cada uno de los parámetros acústicos extraídos de las muestras de voz para generar los vectores de tamaño fijo de cada una de las características.

Los GMM tratan de encontrar un conjunto o mezcla (*mixture*) de distribuciones de probabilidad *gaussianas* como una combinación lineal de las mismas, que permitan modelar un conjunto de datos de entrada. Las distribuciones de probabilidad se obtienen de los vectores de características de las muestras de voz, que se combinarán para generar los modelos GMM correspondientes.

Una distribución *gaussiana* en una dimensión viene dada por la ecuación 2.7:

$$p(x|\theta_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\Big(\frac{x-\mu_i}{\sigma_i}\Big)^2\Big) \tag{2.7}$$

donde:

μ_i es la media de la distribución

σ_i es la varianza

Para crear el modelo, es necesario calcular los parámetros propios de la distribución *gaussiana*, la media μ_i y la varianza σ_i, para ajustar el modelo a los datos de entrada.

Si se representa una distribución *gaussiana* en n dimensiones, la expresión viene dada por la ecuación 2.8:

$$p(x|\theta_i) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\Big) \tag{2.8}$$

donde:

μ_i es el vector de medias de las distribuciones de dimensión n

Σ_i es la matriz de covarianzas de tamaño $n \times n$

Los GMM se desarrollaron para modelar distribuciones más complejas, en las que existen correlaciones no lineales o que cuentan con diferentes modas. La expresión general del conjunto de distribuciones *gaussianas* viene dado por la ecuación 2.9:

$$p(x) \approx \sum_k c_k p(x|\theta_k) \tag{2.9}$$

donde:

$\{c_k\}$ es un conjunto de pesos

$\{p(x|\theta_k)\}$ es un conjunto de componentes *gaussianas*

Cada observación será generada por una de las distribuciones *gaussianas*, que será escogida con una probabilidad $c_k = p(\theta_k)$

Para generar los modelos que describen a los datos de entrada, es necesario calcular los parámetros que definen a la distribución a partir de los datos. Existen diferentes métodos para estimar dichos parámetros siendo uno de los más habituales el algoritmo de *Expectation Maximization* (EM).

Algoritmo de Expectation Maximization El algoritmo de EM sirve para calcular los parámetros de máxima verosimilitud de un conjunto de distribuciones a partir de un número conocido a priori de componentes. Este método se utiliza para estimar los parámetros de los modelos cuando no se pueden calcular directamente, aplicando un proceso iterativo basado en:

- Tomar unos valores arbitrarios para resolver un conjunto de ecuaciones.
- Utilizar el resultado obtenido para estimar los valores en el segundo conjunto de ecuaciones.
- Con los resultados obtenidos del segundo conjunto se vuelve al primer conjunto y se ajustan los valores para resolver de nuevo las ecuaciones y encontrar unos parámetros mejores.
- Con los nuevos parámetros se pasa al segundo conjunto y se estiman nuevos parámetros.
- Se repite el proceso hasta que los valores converjan.

El proceso de clasificación se lleva a cabo en dos partes:

- Modelado [Figura 2.3]. En esta fase se construye y entrena el clasificador GMM con los datos de entrenamiento del conjunto, obteniendo los vectores de características de las muestras de voz, las respectivas distribuciones de probabilidad asociadas a los vectores y calculando posteriormente el GMM para cada sub-conjunto.
- Clasificación [Figura 2.4]. Durante esta segunda fase ocurre la propia clasificación de usuarios en base a las muestras de voz. Teniendo definidos los modelos GMM, se alimenta al clasificador con datos de usuarios desconocidos. El clasificador devolverá una puntuación que describirá la probabilidad de pertenencia de la muestra a los distintos sub-conjuntos de los datos modelados con los GMM.

Figura 2.3: Generación de modelos GMM

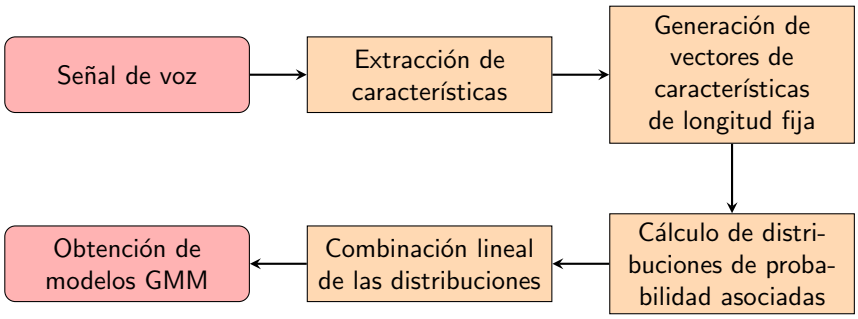
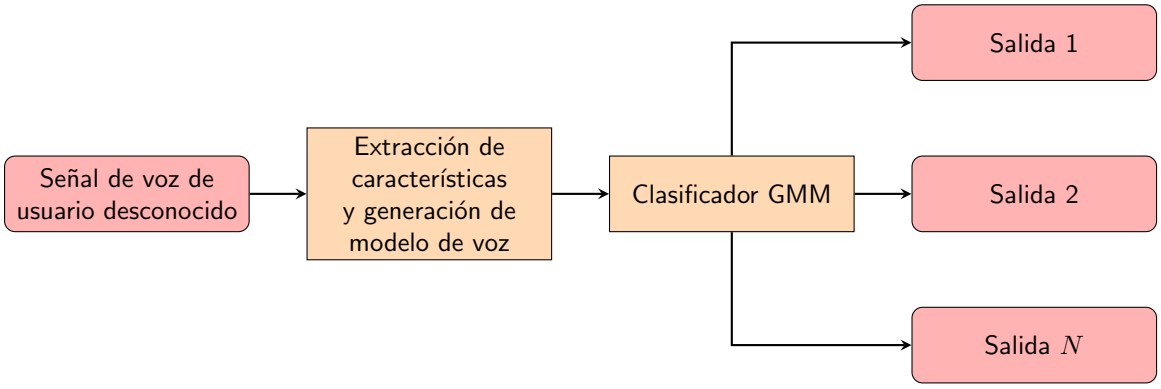


Figura 2.4: Clasificación de muestras con GMM



2.3.2.2. Support Vector Machines

Los SVM son modelos de aprendizaje supervisado que se desarrollaron para llevar a cabo tareas de clasificación y regresión. Se basan en realizar una transformación entre el espacio en el que se representan los datos de entrada y un espacio vectorial alternativo de mayor dimensión, de tal forma que la distancia entre los datos de entrenamiento más cercanos de una categoría y otra (margen) sea mayor en este nuevo espacio, y por tanto el error al clasificar ejemplos en cada una de las dos categorías sea menor. Para separar los datos de entrada, el algoritmo SVM genera hiperplanos en el nuevo espacio vectorial, de tal forma que la separación entre clases ocurra con el menor error posible.

La equivalencia entre el espacio de entrada y el nuevo espacio vectorial de mayor dimensión se ajusta con la función *kernel* de SVM $k(.,.)$. Los hiperplanos en el espacio vectorial de mayor dimensión se definen como el conjunto de puntos cuyo producto escalar con un vector en ese espacio es constante, siendo el conjunto de vectores aquellos ortogonales que definen el hiperplano. Los vectores que conforman el hiperplano se pueden definir como una combinación lineal de parámetros α_i con los vectores de características extraídos de los datos de entrada x_i. El separador óptimo vendrá dado por la suma de las funciones de *kernel* [Ecuación 2.10] [30]:

$$f(x) = \sum_{i}^{L} \alpha_i t_i k(x, x_i) + d \tag{2.10}$$

donde:

L es el número de puntos de entrenamiento

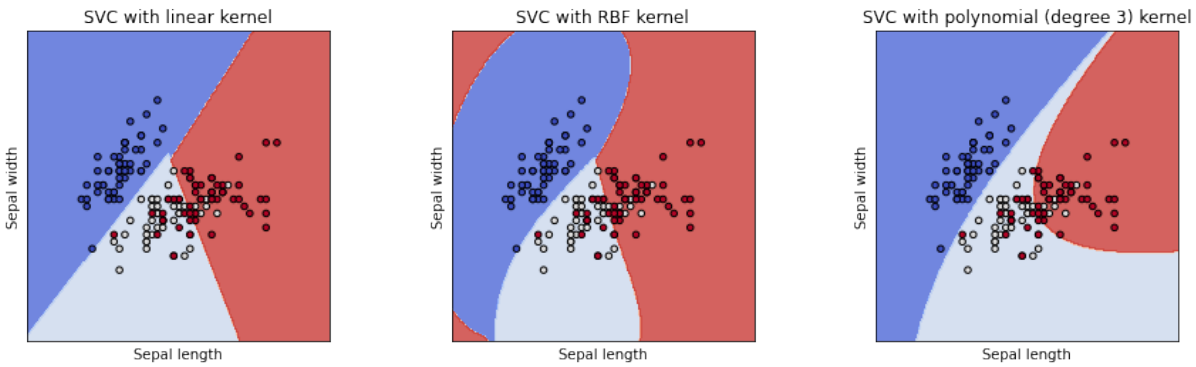
x_i son los vectores obtenidos de los datos de entrenamiento

d es un valor constante

La clasificación se realizará en base al valor que se obtenga de $f(x)$. Si el resultado de $f(x)$ está por encima de valor umbral definido se clasificará como la clase $+1$ y si está por debajo se clasificará como la clase -1.

Funciones de kernel Existen diferentes posibilidades a la hora de configurar las funciones de *kernel* que se van a usar para transformar los datos entre el espacio de entrada y el espacio vectorial de salida En la Figura 2.5 se representa a modo de ejemplo el efecto de distintas funciones de *kernel* en la clasificación sobre el *dataset iris* [33].

Figura 2.5: Ejemplo de clasificación con diferentes funciones de *kernel*



Radial Basis Function kernel Una de las funciones más adecuadas cuando no se tiene conocimiento previo de los datos es la función *gaussiana* o *Radial Basis Function* (RBF), que viene dada por la función 2.11 [34]:

$$k(x, x_i) = \exp(\gamma ||x - x_i||^2) \tag{2.11}$$

donde:

γ es el parámetro de ajuste de la función, que tiene que cumplir $\gamma > 0$

Clasificación con SVMs Como se ha expuesto anteriormente, los SVM son fundamentalmente algoritmos de clasificación binarios, es decir, distinguen entre dos clases, que se han definido como $+1$ o -1 en este caso. Para trabajar con múltiples clases, se recurren a diferentes aproximaciones para transforma el clasificador binario en uno multiclase. Uno de los métodos usados comúnmente es la aproximación conocida como *one-versus-the-rest*, en la cuál se construyen K SVM por separado. El modelo $k - esimo$ se entrena con los datos de la clase C_k como ejemplos positivos, mientras que el resto de $K-1$ clases se utilizan como ejemplos negativos. Sin embargo, esta aproximación presenta una serie de problemas, derivados de las posibles inconsistencias que pueden producirse al usar múltiples clasificadores entrenados con conjuntos diferentes. Además, existe otro problema derivado de la selección de los conjuntos de datos. Al seleccionar una sola clase como ejemplos positivos y el resto como ejemplos negativos, se provoca que el conjunto no sea simétrico, perdiendo la distribución original de los datos. Otra aproximación para el uso de SVM en problemas de clasificación múltiple es la conocida como *one-versus-one*, que consiste en entrenar $K\frac{K-1}{2}$ SVM binarios para todos los pares de clases posibles, clasificando los ejemplos de *test* en función de que clase tenga el número de votos más alto. Este modelo también presenta desventajas, como un proceso de *test* más complejo desde el punto de vista computacional, ya que es necesario evaluar cada punto de *test* con cada uno de los SVM construidos [35].

Los SVM son modelos muy potentes con ventajas [34] como su efectividad en espacios de grandes dimensiones y el uso de puntos del conjunto de entrenamiento para generar la función de decisión, que los hace eficientes en el uso de memoria. Sin embargo, también presentan ciertas desventajas que habrá que tener en cuenta a la hora de implementar y evaluar los modelos, como la elección cuidadosa de parámetros para obtener un buen rendimiento o su complejidad desde el punto de vista computacional, que aumenta rápidamente conforme crecen los vectores de entrenamiento. La complejidad de la implementación de SVM que utiliza la herramienta *Scikit Learn* (ver Sección 2.4.2 para descripción de la herramienta) varía entre $O(n_{caracteristicas} \times n^2_{muestras})$ y $O(n_{caracteristicas} \times n^3_{muestras})$

2.3.2.3. Random Forest

Los modelos *Random Forest* pertenecen a la familia de métodos de *Machine Learning* conocidos como *Ensemble Learning*. Como los SVM, son utilizados para trabajos de regresión y clasificación. *Random Forest* se basa en construir múltiples árboles de decisión durante el entrenamiento. Cada uno de los árboles se crea de forma aleatoria, y se utilizan para clasificar o predecir en función de los datos de entrada. La decisión final de *Random Forest* viene dada por la media de las predicciones de todos los árboles en el caso de regresión, o por la clase que la mayoría de los clasificadores haya seleccionado en el caso de que se esté realizando una clasificación. En la implementación del algoritmo *Random Forest* que ha desarrollado *Scikit Learn* el procedimiento de elección final del conjunto de árboles ha sido modificado ligeramente. En vez de que cada uno de los árboles seleccione una sola clase, se tiene en cuenta la probabilidad de pertenencia del ejemplo a cada una de las clases, y el resultado final se obtiene seleccionando la clase cuya media de predicciones sea más alta.

Cada uno de los árboles que componen el *ensemble* se entrena con una muestra del conjunto de entrenamiento, que se obtiene aplicando una sustracción con repeticiones al *dataset* de entrenamiento. Además, cada uno de los árboles se entrena con un subconjunto de las características totales para evitar la correlación entre los árboles que puede aparecer al tener características más significativas que otras, y que si son seleccionadas en todos los árboles para el entrenamiento aparecerán como predictores importantes en todos ellos obviando el efecto del resto de características.

Los modelos *Random Forest* destacan frente a los árboles de decisión tradicionales gracias a su robustez frente al sobre-entrenamiento. Mientras que un sólo árbol de decisión con una profundidad muy alta puede aprender patrones no deseados en los datos, *Random Forest* ofrece como resultado la media de los resultados de múltiples árboles de decisión con alta profundidad, reduciendo así la varianza del modelo, y por tanto el sobre-entrenamiento.

Random Forest es un algoritmo eficiente, con una complejidad en su implementación en *Scikit Learn* de $O(M \times N \times \log(N))$, donde M es el número de árboles y N es el número de muestras del conjunto de datos. Además, es posible paralelizar las operaciones del algoritmo al ser independientes los cálculos para el entrenamiento de cada uno de los árboles.

2.3.3. Conjuntos de datos

En esta sección se describen los conjuntos de datos utilizados en el proyecto.

2.3.3.1. Common Voice

El conjunto de datos *Common Voice* [36] [Figura 2.6] ha sido desarrollado por *Mozilla*. Cuenta con voces aportadas por la comunidad de *Common Voice*, que realiza tareas de creación de muestras aportando su voz y de validación, revisando que las grabaciones de los contribuyentes corresponden con el mensaje que tienen que leer. El *dataset* se actualiza en intervalos de aproximadamente seis meses, añadiendo nuevas muestras de audio y validando tanto las nuevas como las ya existentes. Existen diferentes versiones del conjunto de datos por idiomas. Para el estudio se hará uso de las versiones en inglés y español, que permitirán evaluar el rendimiento de los clasificadores de voz para distintos idiomas.

Figura 2.6: Imagen del proyecto de *Common Voice*

Common Voice en inglés La versión en inglés es la más extensa, con 56 Gb de información. El conjunto de datos en inglés se encuentra en la versión 6.1 a fecha actual, habiendo sido actualizado por última vez el 11 de Diciembre de 2020. Cuenta con un total de 2181 horas de contenido, de las cuales 1686 horas están validadas. En la grabación han participado 66173 personas con una distribución de acentos, género y edad como se indica en la Tabla 2.1.

Tabla 2.1: Distribución por acento, género y edad del conjunto de datos *Common Voice* en inglés

Propiedad	Distribución
Acento	
United States English	24 %
England	10 %
India and South Asia	5 %
Australian English	3 %
Canadian English	3 %
Scottish English	2 %
Irish English	1 %
Southern African	1 %
New Zealand English	1 %
Edad	
Teens ($<$ 19 años)	6 %
Twenties (19-29 años)	24 %
Thirties (30-39 años)	14 %
Fourties (40-49 años)	10 %
Fifties (50-59 años)	4 %
Sixties (60-69 años)	4 %
Seventies (70-79 años)	1 %
Eighties (80-89 años)	$<$ 1 %
Nineties ($>$ 90 años)	$<$ 1 %
Género	
Male	47 %
Female	15 %

Common Voice en español Para el conjunto de datos en español se ha escogido la versión 2, que se publicó el 26 de Junio de 2019. Tiene un tamaño de 823 Mb y cuenta con 31 horas de contenido, de las cuales 26 han sido validadas. En el *dataset* se recogen 602 voces diferentes, con una distribución de acentos, género y edad indicada en la Tabla 2.2.

Tabla 2.2: Distribución por acento, género y edad del conjunto de datos *Common Voice* en español

Propiedad	Distribución
Acento	
España: Norte peninsular	29 %
Andino-Pacífico	19 %
Chileno	8 %
Rioplatense	7 %
España: Centro-sur peninsular	4 %
México	3 %
España: Sur peninsular	3 %
España: Islas Canarias	2 %
Español de Filipinas	1 %
Caribe	1 %
América Central	1 %
Edad	
Teens ($<$ 19 años)	2 %
Twenties (19-29 años)	32 %
Thirties (30-39 años)	12 %
Fourties (40-49 años)	16 %
Fifties (50-59 años)	17 %
Sixties (60-69 años)	2 %
Seventies (70-79 años)	$<$ 1 %
Eighties (80-89 años)	$<$ 1 %
Nineties ($>$ 90 años)	$<$ 1 %
Género	
Male	73 %
Female	9 %

Cada conjunto de datos se distribuye en un archivo comprimido, que sigue una estructura similar a la indicada en el fragmento a continuación.

Fragmento de código 2.1: Estructura de los archivos del conjunto de datos *Common Voice*

```
[lang].tar.gz/
|-- clips/
|   |-- *.mp3 files
|-- dev.tsv
|-- invalidated.tsv
|-- other.tsv
|-- test.tsv
|-- train.tsv
|-- validated.tsv
```

El nombre del archivo viene dado por la abreviación del idioma correspondiente a las grabaciones según los códigos de la norma ISO 639-1. Como ejemplo, los nombres del conjunto de datos en inglés y en español corresponden a `en.tar.gz` y `es.tar.gz` respectivamente.

Cada una de las grabaciones dispone de un conjunto de características que contienen un identificador del usuario que la grabó, datos de validación e información demográfica. Los campos que se incluyen para cada una de las grabaciones son:

- `client_id`. *Hash* identificador del usuario que realizó la grabación de la muestra.
- `path`. *Path* relativo a la muestra de audio
- `up_votes`. Número de personas que han dicho que el audio coincide con el texto.
- `down_votes`. Número de personas que han dicho que el audio no coincide con el texto.
- `age`. Edad de la persona que grabó el audio.
- `gender`. Género de la persona que grabó el audio.
- `accent`. Acento de la persona que grabó el audio.
- `segment`. Indica si la muestra pertenece a un conjunto de datos concreto.

Como se ha mostrado en el Fragmento 2.1, las grabaciones se encuentran en el directorio `clips/`, en formato *.mp3*, y las descripciones de las muestras de audio se adjuntan en archivos *.tsv*, separador por categorías:

- `invalidated.tsv`. Conjunto de muestras de audio que han recibido más de dos valoraciones y cuyo número de votos negativos es mayor que el número de votos positivos (`down_votes` $>$ `up_votes`) o que han recibido tres valoraciones o más y que tienen igual número de votos positivos que negativos (`down_votes` $=$ `up_votes`).
- `validated.tsv`. Conjunto de muestras de audio con más de dos valoraciones en las que el número de votos positivos supera al de negativos (`down_votes` $<$ `up_votes`).
- `other.tsv`. Conjunto de muestras que no han recibido suficientes valoraciones para entrar en uno de los grupos anteriores.
- `train.tsv`. Conjunto de muestras dividido para entrenamiento
- `test.tsv`. Conjunto de muestras divididas para *test*.
- `dev.tsv`. Conjunto de muestras divididas para validación.

2.4. Herramientas

Para la realización del clasificador se ha utilizado el lenguaje de programación *Python*, debido a la gran cantidad de librerías y utilidades disponibles y su facilidad de uso. En cuanto a las herramientas específicas utilizadas en el proyecto, se ha recurrido a la librería *openSMILE* para la extracción de características de las muestras de audio y la librería *Scikit Learn* para el desarrollo de los modelos basados en Inteligencia Artificial.

2.4.1. openSMILE

open-source Speech and Music Interpretation by Large-space Extraction (*openSMILE*) [37] [Figura 2.7] es una herramienta de código abierto para la extracción de características de señales de audio que se utilizará para generar los vectores de características de cada muestra de voz con los que se construirá el conjunto de datos que servirá para realizar el proceso de clasificación posterior. La herramienta está desarrollada en *C++* y ofrece una *API* para *Python* que permite un uso más directo de la herramienta a través de conjuntos predefinidos de características.

Figura 2.7: OpenSMILE

2.4.2. Scikit Learn

Scikit Learn [38] [Figura 2.8] es una librería escrita en *Python* que incorpora funcionalidades y herramientas del campo de *Machine Learning*. En la librería se incluyen todos los algoritmos de utilidad para el proyecto, así como herramientas adicionales que facilitan la creación de conjuntos de entrenamiento y validación, evaluación de los modelos, etc.

Figura 2.8: Scikit Learn

Capítulo 3

Análisis de dispositivos comerciales

En este capítulo se detalla el análisis de seguridad y privacidad realizado sobre los asistentes personales comerciales más utilizados, *Siri* de *Apple*, *Assistant* de *Google* y *Alexa* de *Amazon*.

En la Sección 3.1 se describen las pruebas que se han realizados en los dispositivos y las secciones siguientes se dedican al estudio de cada uno de los dispositivos. Como se puede ver en la descripción de las pruebas en la siguiente sección, el análisis se ha centrado en la seguridad de los dispositivos durante las diferentes etapas de interacción que un usuario puede tener con él, para detectar problemas de seguridad que han permitido enfocar el desarrollo de la herramienta de clasificación [Capítulo 4].

3.1. Descripción de las pruebas

Las pruebas de características se han dividido en cuatro categorías, que cubren las diferentes etapas de interacción de un usuario con el dispositivo y los posibles ajustes de seguridad y privacidad que pueda hacer. Con estos ámbitos se ha tratado de obtener una visión global acerca de la seguridad y privacidad en los SPA, centrando los aspectos del estudio en configuraciones y usos habituales que una persona puede hacer de los dispositivos.

En la siguiente lista se muestra la descripción de las pruebas que se han realizado para cada uno de los dispositivos:

- Instalación. Pruebas que cubren tanto la instalación del propio dispositivo con el SPA como el registro de nuevos accesorios y *skills* de terceros.
    - Proceso de instalación del SPA. Se estudia como se lleva a cabo la instalación del asistente personal, que dispositivos y aplicaciones adicionales son necesarios y que configuraciones pueden realizarse.
    - Instalación de nuevos dispositivos conectados. Se comprueba el proceso de instalación de accesorios conectados al asistente y que permisos o configuraciones se pueden establecer para restringir su uso.
    - Instalación de *skills* de terceros. Se analiza el proceso de instalación de *skills* o desarrollos de terceros para ser invocados desde el asistente, si cuentan con esta funcionalidad.
- Interacción. Pruebas que cubren las posibles opciones que un usuario tiene para controlar la interacción que se hace con el SPA, con dispositivos conectados y *skills* de terceros.
    - Interacción con el SPA y dispositivos conectados. Se comprueba como un usuario puede controlar el uso que se hace del asistente y los accesorios conectados.
    - Interacción con *skills* de terceros. Se analiza si existen controles que permitan definir perfiles de uso para las *skills* de terceros incorporadas al asistente.

- Funcionalidad. Pruebas que cubren opciones que permitan controlar las funcionalidades del asistente, como pagos o reproducción de contenido.
    - Pagos y transacciones. Se estudia la posibilidad de realizar pagos desde el asistente, comprobando las configuraciones y restricciones que un usuario pueda establecer.
    - Posibilidad de crear perfiles "seguros" para menores. Se comprueba si los asistentes cuentan con opciones para crear perfiles restringidos para menores de forma sencilla, que permitan controlar el contenido multimedia que se reproduce desde el asistente, los accesorios con los que pueda interactuar, pagos que se puedan realizar, etc.
- Privacidad y seguridad. Esta categoría recoge las pruebas que evalúan las funciones de seguridad con las que cuenta el asistente, así como las posibilidades que tiene un usuario para controlar el uso que se hace de su información.
    - Control de respuestas que contengan información personal. Los asistentes pueden incluir información personal de los usuarios como respuestas a las peticiones. En esta pruebas se evalúa la capacidad de un usuario de controlar el uso de su información.
    - Métodos de autenticación. Se analizan los métodos de autenticación de los que dispone el asistente, así como su efectividad real.
    - Filtrado de voz no humana. Se comprueba si el asistente es susceptible a ser activado por voces de origen sintético o artificial, como un mensaje grabado o un sistema *Text To Speech*.
    - Interacción con el historial de conversaciones. Se estudian las opciones que se proporcionan al usuario del asistente para controlar y visualizar el historial de conversaciones.

3.2. Apple HomePod Mini (2020)

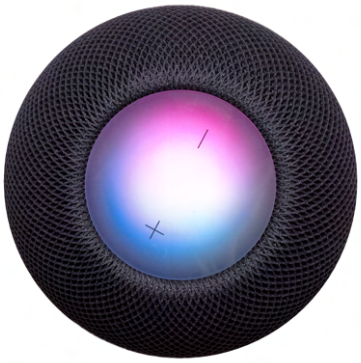
Apple es la tercera marca en el mercado de los altavoces inteligentes más utilizada en el mundo si se omiten aquellas que sólo operan y ofrecen sus productos en el mercado asiático. Si bien su cuota de mercado estimada es más reducida que sus competidores, un 6 % frente al 31.4 % de *Google Assistant* y el 31.7 % de *Amazon Alexa* [6], su SPA, *Siri*, es el asistente inteligente que está instalado en el mayor número de dispositivos en todo el mundo, con 500 millones de dispositivos que lo usan [17]. La gran presencia de *Siri* en el mercado viene impulsado por la popularidad de los *smartphones* de la marca [39] o sus *tablets* [40], frente a un *Amazon* que centra su oferta de *hardware* de consumo en el sector de los altavoces inteligentes y la integración de su asistente en accesorios de terceros, pero que sin embargo, no dispone de alternativas propias que puedan competir con los *iPhones* o *iPads* de *Apple*. *Siri* se diferencia del resto de SPA en la integración con desarrollos de terceros. Mientras que *Alexa* permite la introducción de *skills* para ser invocadas desde los dispositivos y *Google Assistant* proporciona una funcionalidad similar a través de las *actions*, el ecosistema de *Siri* es más cerrado, y no permite la invocación de funcionalidades de terceros que se ejecutan en servidores externos a los del fabricante.

Apple ha comercializado hasta la fecha dos modelos de altavoces inteligentes con su SPA integrado en la familia *HomePod*. En 2018, se introdujo el primer modelo, de nombre *HomePod*, que dio comienzo a la línea de altavoces inteligentes. La marca mantuvo su fabricación hasta el 12 de Marzo de 2021, momento en el que se discontinuó, para centrarse en el nuevo modelo, el *HomePod Mini*, lanzado en Noviembre de 2020.

El *HomePod Mini* cuenta con control táctil integrado en la parte superior del dispositivo, que sirve para controlar la reproducción de contenido e invocar al asistente inteligente, además de disponer por el control por voz clásico para activar al asistente. No dispone de controles adicionales para apagar los micrófonos del dispositivo.

Figura 3.1: *HomePod Mini*

Figura 3.2: Vista superior del *HomePod Mini*



Para poder dar de alta y usar un *HomePod* es necesario contar con una cuenta de *iCloud* y un dispositivo móvil de la marca, como un *iPhone* o un *iPad* [Tabla 3.1 [41]].

Tabla 3.1: Dispositivos compatibles con el *HomePod Mini*

Dispositivo	Versión
iPhone[1]	*SE*
	6s[2]
iPod Touch[1]	7ª generación[2]
	Pro
iPad[3]	5ª generación[2]
	Air 2[2]
	Mini 4[2]

[1] Los dispositivos deben ejecutar la última versión del sistema operativo (*iOS 14* a fecha actual)
[2] O versiones posteriores
[3] Los dispositivos deben ejecutar la última versión del sistema operativo (*iPadOS14* a fecha actual)

El asistente de *Apple* se controla desde la aplicación *Casa*. Desde ella se añaden nuevos dispositivos, se crean automatizaciones y se configuran tanto el asistente como los accesorios.

Figura 3.3: Aplicación *Casa* en la *App Store*

Según el análisis de privacidad que el desarrollador (*Apple* en este caso) ha publicado para la aplicación, se recogen identificadores del dispositivo para mejorar la funcionalidad de la aplicación, que están vinculados con el usuario, y datos de uso y diagnóstico, que no están vinculados con el usuario.

El usuario que instala el *HomePod Mini* con su dispositivo móvil se convierte en el administrador del SPA y del ecosistema completo de la casa, pudiendo dar de alta a más usuarios, añadir dispositivos o eliminarlos, definir configuraciones para los mismos, etc.

Las pruebas se han realizado enlazando el *HomePod Mini* a un *iPhone X* (2017) con versión de *iOS* 14.6.

A continuación se detallarán los resultados de cada una de las pruebas definidas en el *HomePod Mini*. Al final de la sección, en la Tabla 3.2, se incluye un resumen con los resultados de las pruebas.

3.2.1. Pruebas de instalación

3.2.1.1. Instalación del SPA

El *HomePod Mini* se instala escaneando el patrón que aparece en el control táctil en el momento del emparejamiento con la cámara del dispositivo móvil con el que se registra.

3.2.1.2. Instalación y edición de nuevos dispositivos conectados

Se permite controlar de forma individualizada que usuarios pueden dar de alta nuevos accesorios y editar los ya existentes. El acceso al control de dispositivos se encuentra en la sección del perfil de usuario registrado en la casa, al que se accede como se muestra en las Figuras 3.4, 3.5, 3.6 y 3.7.

Figura 3.4: Acceso a ajustes

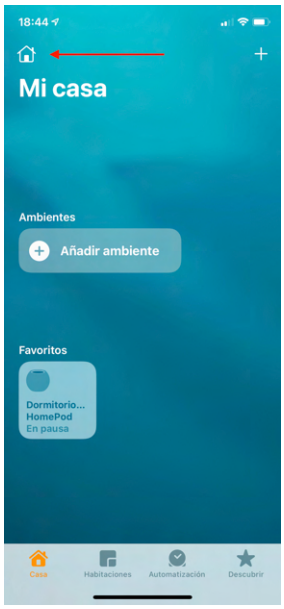


Figura 3.5: Acceso a ajustes de la casa

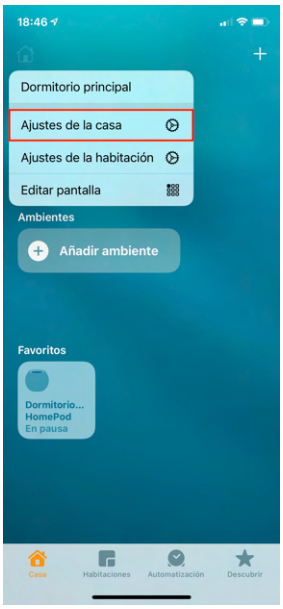


Figura 3.6: Acceso al perfil del usuario

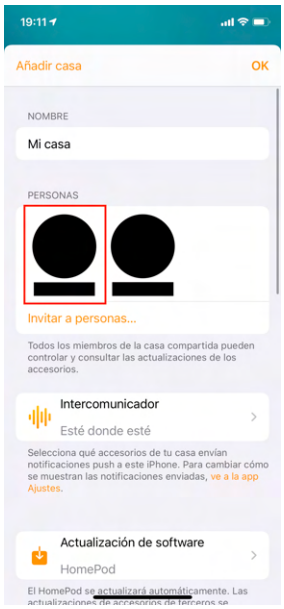
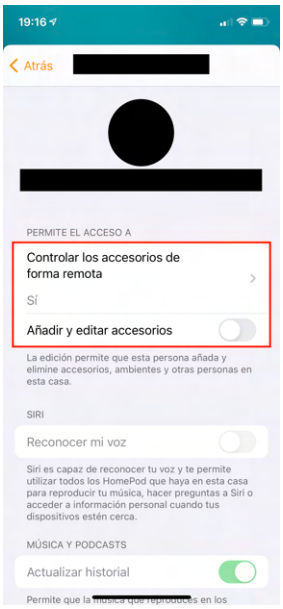


Figura 3.7: Acceso a opciones de control



3.2.1.3. Instalación de skills de terceros

En el *HomePod Mini* no se permite añadir desarrollos de terceros para invocarlos con el asistente.

3.2.2. Pruebas de interacción

3.2.2.1. Interacción con el SPA y dispositivos conectados

Como se mostró en la introducción del dispositivo al comienzo de la sección, un usuario puede interactuar con el *HomePod Mini* a través del control por voz, invocando al asistente a través del comando de activación, o mediante el control táctil en la parte superior del dispositivo. Ambas opciones son configurables desde la aplicación *Casa*, como se muestra en las Figuras 3.8 y 3.9.

Figura 3.8: Acceso a ajustes del *HomePod Mini*

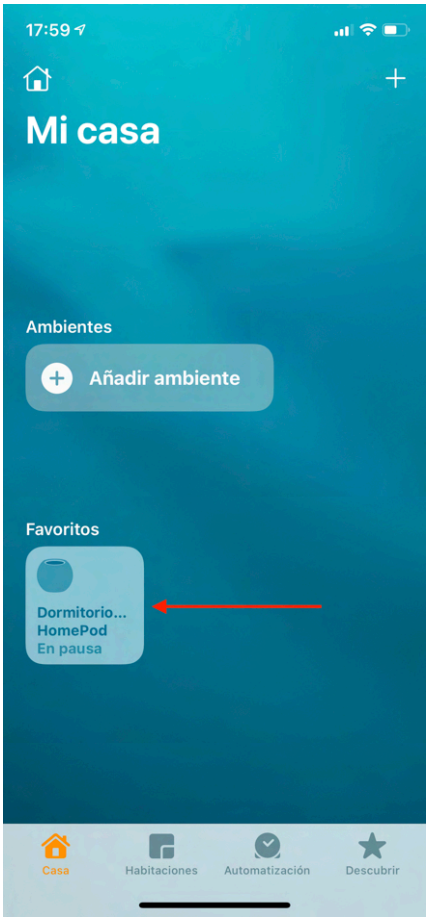


Figura 3.9: Ajustes de interacción con el SPA

Con respecto al control de acceso a dispositivos, se puede realizar a dos niveles, uno relacionado con la reproducción de contenido multimedia y otro relativo al control de dispositivos conectados al asistente.

Control de dispositivos multimedia La configuración se realiza en el apartado de ajustes de la aplicación *Casa*, desde donde se accede al control de acceso a los dispositivos de reproducción multimedia. En las Figuras 3.10, 3.11, 3.12, 3.13 se muestra el proceso para acceder a los permisos de reproducción de contenido en altavoces y televisores. En la Figura 3.13 se muestran las tres opciones con las que cuenta un usuario para configurar los permisos de uso:

- `Todos`. Cualquier usuario puede ver y reproducir contenido en los dispositivos. Adicionalmente se ofrece una opción de establecer una contraseña para usar los dispositivos.
- `Cualquiera en la misma red`. Cualquier usuario que esté conectado a la misma red que los dispositivos multimedia puede enviar contenido a los mismos. También se ofrece una opción adicional de establecer una contraseña para usar los dispositivos.
- `Solo quienes comparten esta casa`. Como su nombre indica, solo los usuarios que el administrados haya dado de alta en la aplicación podrán hacer uso de los dispositivos multimedia para reproducir contenido, sin importar desde donde accedan al los dispositivos.

Figura 3.10: Acceso a ajustes

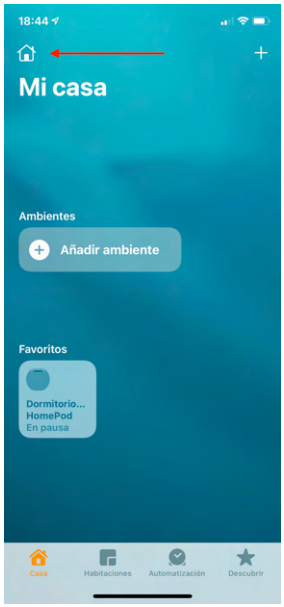


Figura 3.11: Acceso a ajustes de la casa

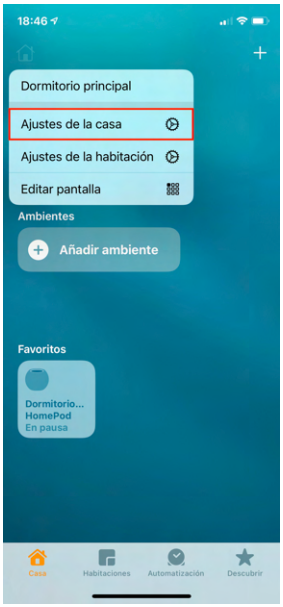
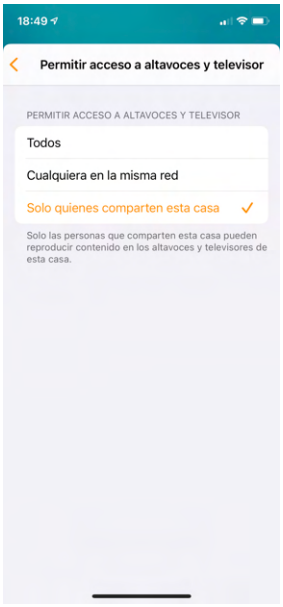


Figura 3.12: Acceso a control de dispositivos

Figura 3.13: Opciones de control



Control de dispositivos conectados El control de uso de dispositivos conectados se realiza para cada usuario registrado en la casa por el administrador. Siguiendo el flujo de las Figuras 3.14 y 3.15, se accede al perfil del usuario como se muestra en las Figuras 3.16 y 3.17. Desde este punto, como se observa en la Figura 3.18, se puede dar acceso a un usuario para que controle de forma remota los accesorios conectados al asistente.

Figura 3.14: Acceso a ajustes

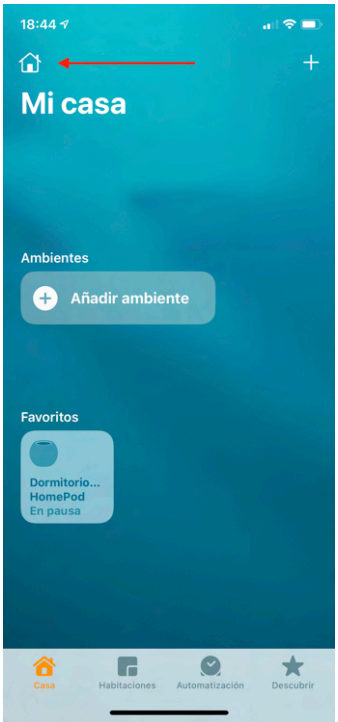


Figura 3.15: Acceso a ajustes de la casa

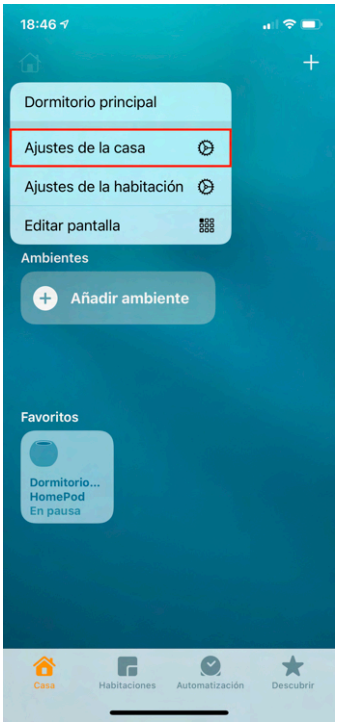


Figura 3.16: Acceso al perfil del usuario

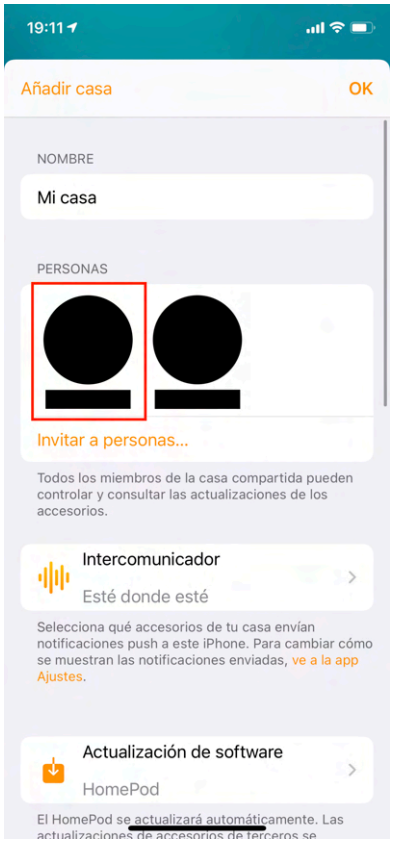


Figura 3.17: Acceso a control de dispositivos

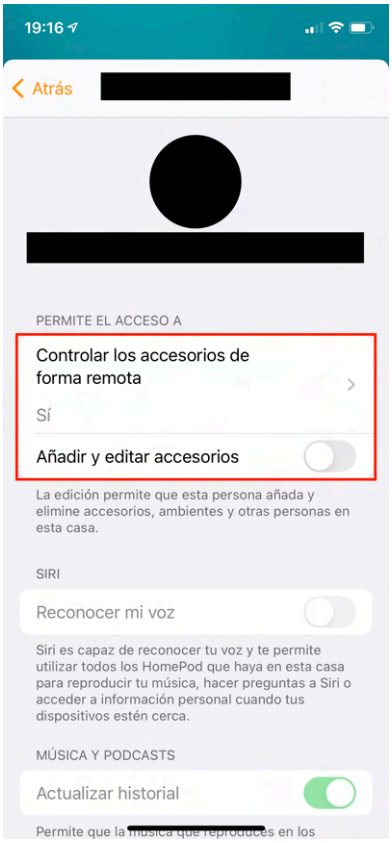
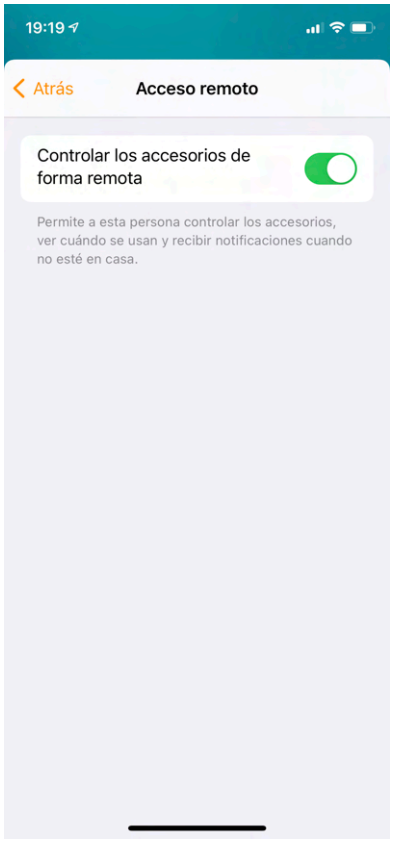


Figura 3.18: Opciones de control de dispositivos



3.2.2.2. Interacción con skills de terceros

Como se ha expuesto en la Sección 3.2.1.3, no se pueden configurar ni utilizar *skills* de terceros en el *HomePod Mini*.

3.2.3. Pruebas de funcionalidad

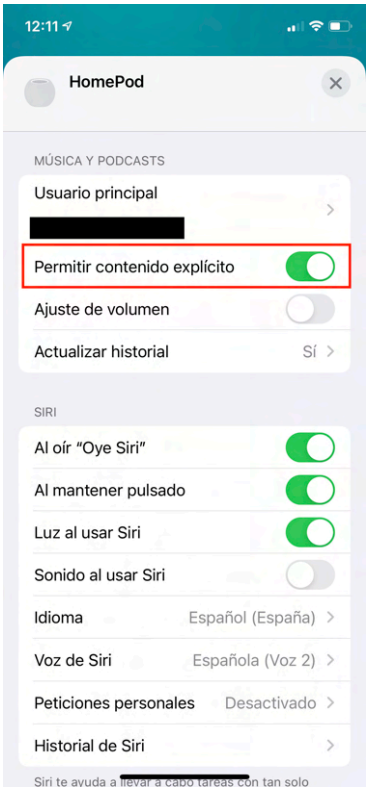
3.2.3.1. Pagos y transacciones

El *HomePod Mini* no permite realizar pagos ni compras desde el asistente.

3.2.3.2. Posibilidad de crear perfiles "seguros" para menores

No existe una opción dedicada para crear usuarios a los que sólo se les permita hacer un uso restringido del asistente, siendo necesario usar los controles genéricos en el perfil del usuario una vez se crea, para desactivar la posibilidad de añadir, editar y controlar accesorios [Figura 3.17]. En cuanto a la reproducción de contenido multimedia en el altavoz, no existe la posibilidad de definir un control por usuario que permita restringir la reproducción de contenido explícito. La posibilidad que se ofrece es desactivar la reproducción de contenido explícito por dispositivo para todos los usuarios, como se muestra en la Figura 3.19.

Figura 3.19: Control de contenido explícito



3.2.4. Pruebas de privacidad y seguridad

3.2.4.1. Control de respuestas que contengan datos personales

El *HomePod Mini* es capaz de distinguir usuarios (funcionalidad no disponible en español actualmente), permitiendo a cada persona interactuar con sus recordatorios, agenda de contactos, mensajes, etc. Para realizar preguntas que incluyen información personal en su respuesta, se lleva a cabo un proceso de detección de presencia del usuario, para comprobar que está en casa y evitar que se obtenga información del asistente sin su conocimiento. La detección de presencia se realiza por medio del dispositivo móvil principal que el usuario tenga registrado en su cuenta de *iCloud*, comprobando que esté cerca del *HomePod Mini*. La configuración de las peticiones personales se realiza desde los ajustes del dispositivo, como se muestra en las Figuras 3.20, 3.21 y 3.22. Si bien es necesario el reconocimiento de voz para ofrecer respuestas personales en un entorno multiusuario, es posible activar las respuestas personales para la cuenta principal en el *HomePod Mini*, bajo advertencia previa de que cualquier usuario podría acceder a los datos personales, ya que no se está realizando ningún tipo de control del usuario que emite las preguntas, como se muestra en la figura.

Figura 3.20: Acceso a ajustes del *HomePod Mini*

Figura 3.21: Acceso a peticiones personales

Figura 3.22: Control de peticiones personales

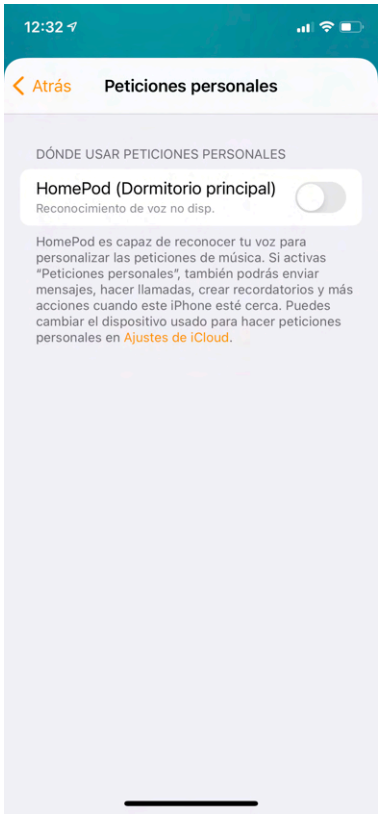
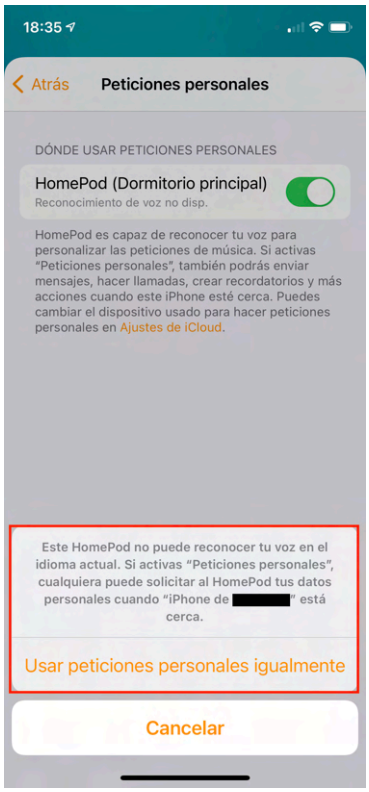


Figura 3.23: Advertencia al activar peticiones personales sin reconocimiento de voz



3.2.4.2. Métodos de autenticación

El *HomePod Mini* proporciona autenticación a los usuarios a través del reconocimiento de voz y la detección de presencia del dispositivos principal del usuario.

3.2.4.3. Filtrado de voz no humana (voz de usuario pregrabada y sistemas Text To Speech)

Se ha comprobado que el dispositivo no realiza un filtrado de los mensajes de activación que le llegan antes de procesarlos. Se han analizado dos alternativas, un mensaje pregrabado con la voz del usuario y un sistema *Text To Speech* (TTS), y en ambos escenarios el asistente responde a la frase de activación permitiendo interactuar con él, tal como lo haría un usuario.

3.2.4.4. Posibilidad de interactuar con el historial de conversaciones con el asistente

Desde la aplicación *Casa* se puede acceder a los ajustes del historial de conversaciones que se han tenido con el asistente desde el *HomePod Mini*. Desde esta sección se da la opción de enviar una petición al fabricante para que el historial se elimine de sus servidores. Sin embargo, no existe una opción para visualizar el historial de conversaciones con el asistente. Entrando a los ajustes del dispositivo como se indica en la Figura 3.24, se accede a los ajustes del historial como aparece en las Figuras 3.25 y 3.26, donde es posible eliminar el contenido de las conversaciones con el SPA.

Figura 3.24: Acceso a ajustes del *HomePod Mini*

Figura 3.25: Acceso al historial de conversaciones

Figura 3.26: Borrado del historial de conversaciones

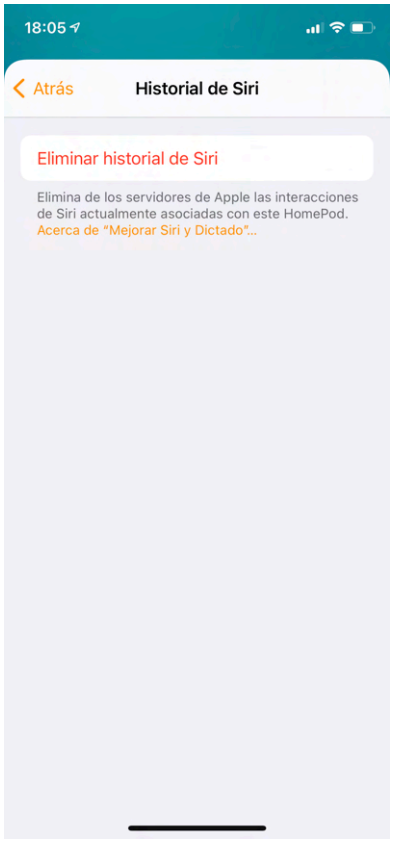


Tabla 3.2: Resumen de resultados de pruebas en el *HomePod Mini*

Característica	Resultado obtenido
Instalación	
Proceso de instalación del SPA	Es necesaria una cuenta de *iCloud* y un dispositivo de *Apple*. Los datos de *Wi-Fi*, *Siri* y preferencias se comparten desde el *iPhone*
Instalación de nuevos dispositivos conectados	Se permite configurar si un usuario tiene permiso para añadir o editar dispositivos
Instalación de *skills* de terceros	No se pueden configurar *skills* de terceros
Interacción	
Interacción con el SPA y dispositivos conectados	Se puede desactivar la invocación por voz o el control táctil. Se permite desactivar el uso de dispositivos multimedia y particularizar por usuario el control de accesorios conectados
Interacción con skills de terceros	No hay posibilidad de interactuar con *skills* de terceros
Funcionalidad	
Pagos y transacciones	No se permiten los pagos o compras desde el HomePod
Posibilidad de crear perfiles "seguros" para menores	No hay posibilidad de crear usuarios para menores, es necesario acceder a cada control y desactivarlo de forma manual
Privacidad y seguridad	
Respuestas que contengan información personal	Es necesario activar el reconocimiento de voz para dar respuestas con datos personales. No disponible en español. Se permiten activar las respuestas personales sin reconocimiento de voz bajo advertencia de que cualquier usuario podría acceder a la información
Métodos de autenticación	Autenticación por reconocimiento de voz (no disponible en español)
Filtrado de voz no humana (voz de usuario pregrabada, sistemas TTS)	Un mensaje de activación pregrabado o leído con un sistema TTS puede invocar al asistente
Interacción con el historial de conversaciones	Se permite enviar una petición para eliminar el historial de los servidores. No se puede visualizar el historial de conversaciones

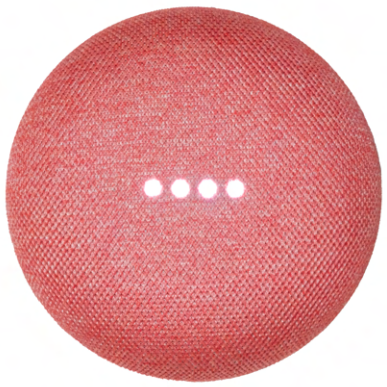
3.3. Google Home Mini (2017)

Con una cuota de mercado estimada del 31.4 % [16] en 2019, *Google* ofrece diferentes dispositivos con su asistente personal incorporado, así como una gama de accesorios inteligentes como cámaras de seguridad y sensores que complementan su ecosistema, además de ofrecer compatibilidad con gran cantidad de fabricantes externos para que incorporen sus productos y los usuarios puedan interactuar con ellos a través de *Assistant*. Todos los dispositivos cuentan con micrófonos incorporados y el modelo *Nest Hub Max* también incluyen cámara [42]. Disponen además de modelos con pantallas y capacidad de interacción táctil con el contenido (*Nest Hub* y *Nest Hub Max*). *Assistant* ofrece la posibilidad de incorporar funcionalidades de otros desarrolladores o *actions* para ser invocadas desde el dispositivo como si se trataran de una funcionalidad nativa [43].

El modelo analizado en este estudio es el *Google Home Mini* [Figuras 3.27 y 3.28], que salió a la venta en Octubre de 2017. El dispositivo cuenta con controles táctil les en la parte superior del dispositivo, un botón dedicado para apagar y encender los micrófonos y luces indicadoras en la parte superior para informar al usuario del estado del dispositivo, por ejemplo, para saber cuando se ha invocado al asistente o cuando se está buscando información acerca de una pregunta que se ha hecho.

Figura 3.27: *Google Home Mini*

Figura 3.28: Vista superior del *Google Home Mini*



Las pruebas se han realizado enlazando el *Google Home Mini* con un *iPhone X* con versión del sistema operativo *iOS 14.6*.

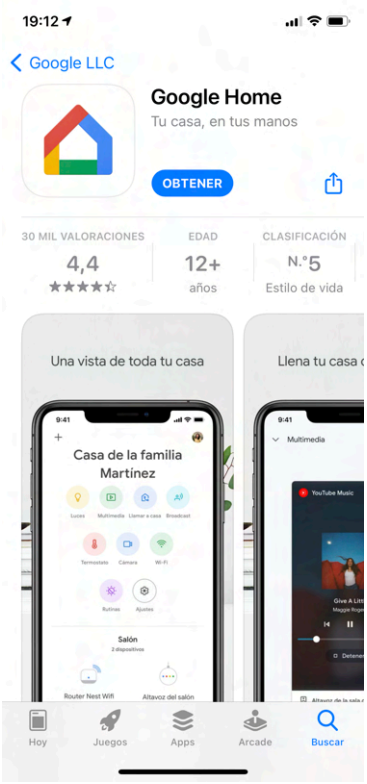
Todas las pruebas se detallan en los apartados siguientes, incluyendo un resumen con los resultados en la Tabla 3.3.

3.3.1. Pruebas de instalación

3.3.1.1. Instalación del SPA

Para instalar y configurar el dispositivo es necesario descargar la aplicación *Google Home* en el *smartphone* [Figura 3.29], siendo la última versión disponible a fecha actual la 2.39.104.

Figura 3.29: Aplicación *Google Home* en la *App Store*



En la sección de privacidad que el desarrollador (*Google*) publica para la aplicación *Google Home*, se puede observar una recolección de datos mucho más significativa si se compara con la aplicación *Casa* del *HomePod Mini*. Por funcionalidad, los datos que se recogen del usuario con la aplicación son:

- Datos vinculados con el usuario.
  - Publicidad de terceros.
    - Historial de navegación.
  - Publicidad o marketing del desarrollador.
    - Datos de contacto, como dirección física y correo electrónico.
    - Identificadores del usuario.
    - Datos de uso de la aplicación.
  - Análisis de datos.
    - Ubicación exacta y/o aproximada.
    - Datos de contacto, como correo electrónico o número de teléfono.
    - Contactos.
    - Contenido del usuario, datos de audio, atención al cliente y otros contenidos.
    - Historial de búsquedas.
    - Identificadores del usuario y/o del dispositivo.
    - Datos de uso de la aplicación.
    - Diagnóstico.
    - Otros datos.
  - Personalización del producto.
    - Ubicación exacta.
    - Datos de contacto, como la dirección de correo electrónico.
    - Contactos.
    - Contenido del usuario.

        - Historial de búsqueda.
        - Identificadores del dispositivo y/o del usuario.
        - Datos de uso.
    - Funcionalidad de la aplicación
        - Historial de compras.
        - Información financiera, como información de pagos.
        - Ubicación exacta y/o aproximada.
        - Datos de contacto, como dirección física, correo electrónico, nombre y número de teléfono.
        - Contactos.
        - Contenido del usuario, como fotos, vídeos, datos de audio y atención al cliente.
        - Historial de búsquedas.
        - Historial de navegación.
        - Identificadores del usuario y/o del dispositivo.
        - Datos de uso.
        - Diagnóstico.
        - Otros datos.

- Datos no vinculados con el usuario.
    - Análisis de datos.
        - Diagnósticos, como datos de errores.
    - Funcionalidad de la aplicación.
        - Diagnósticos, como datos de errores.

Para utilizar la aplicación e instalar el dispositivo es necesaria una cuenta de *Google*. Una vez se inicia sesión en la aplicación, es necesario crear una nueva casa para añadir dispositivos [Figura 3.30].

Es necesario introducir una dirección física para la casa y activar el acceso a la ubicación exacta del *smartphone*, y a la red local para detectar dispositivos a través de *Wi-Fi* y *Bluetooth*. Cuando se establecen los ajustes de la casa, la aplicación detecta automáticamente el *Google Home Mini* [Figura 3.31].

Figura 3.30: Creación de una nueva casa en la aplicación *Google Home*

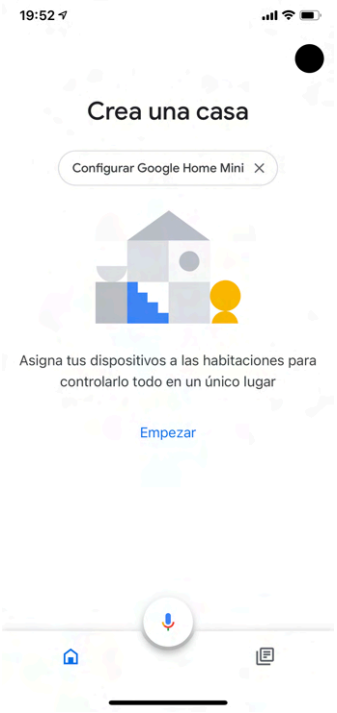
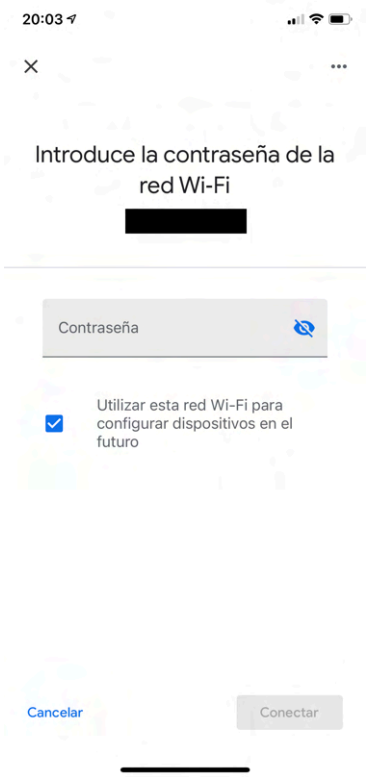


Figura 3.31: Detección del *Google Home Mini* en la aplicación

Cuando se conecta con el dispositivo, se reproduce un sonido para saber si se está conectando al *Google Home Mini* correcto.

El siguiente paso consiste en configurar la conexión a la red *Wi-Fi* introduciendo la contraseña de la red en la aplicación [Figura 3.32].

Figura 3.32: Conexión a la red *Wi-Fi*



Una vez se ha conectado el *Google Home Mini* a la red, se configura el reconocimiento de voz para distinguir a distintos usuarios cuando interactúen con el asistente [Figura 3.33].

Para configurar *Voice Match* se leen diferentes frases para grabar al usuario y generar un modelo con su voz. Aunque se indique que el asistente puede distinguir a varias personas, se muestra un mensaje de advertencia indicando que voces similares o una grabación pueden confundir al asistente [Figura 3.34].

Figura 3.33: Activación de *Voice Match*

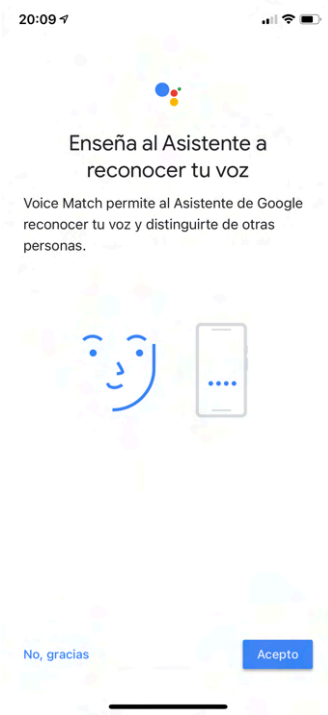
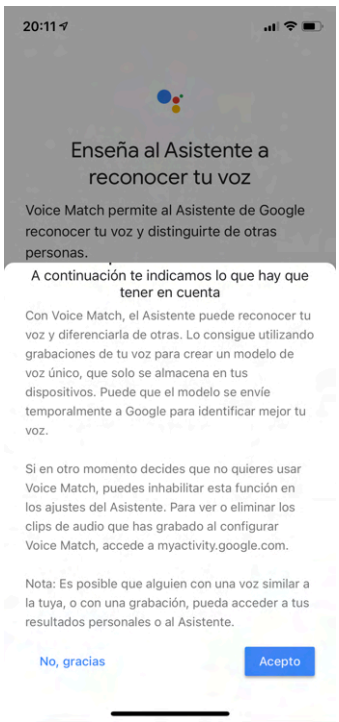


Figura 3.34: Notas sobre el uso de *Voice Match*



Si se configura *Voice Match*, se da la opción al usuario activar los resultados personalizados, que permitirán al asistente interactuar e informar con el calendario, recordatorios o contactos de un usuario específico cuando se reconozca su voz [Figura 3.35].

Figura 3.35: Activación de resultados personalizados

El usuario administrador de la casa puede invitar a otros usuarios a formar parte de ella, lo que les proporcionará acceso total al control de dispositivos, usuarios y ajustes de la casa [Figuras 3.36, 3.37, 3.38 y 3.39].

Figura 3.36: Acceso y control que se comparten con invitados

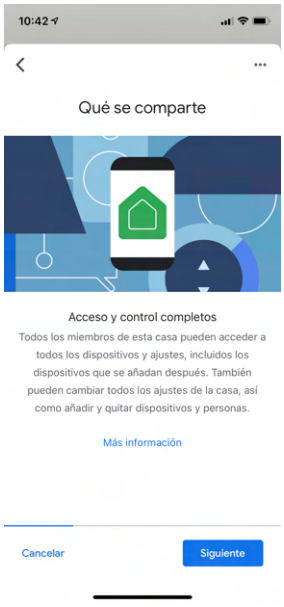


Figura 3.37: Actividad que se comparte con invitados

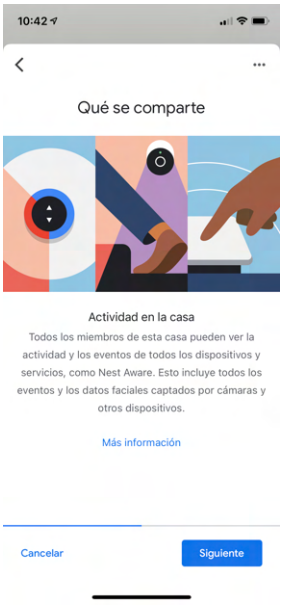


Figura 3.38: Información de contacto que se comparte con invitados

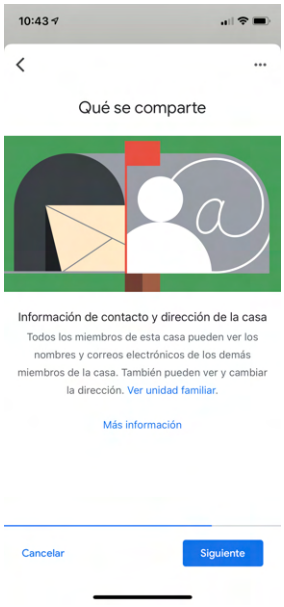
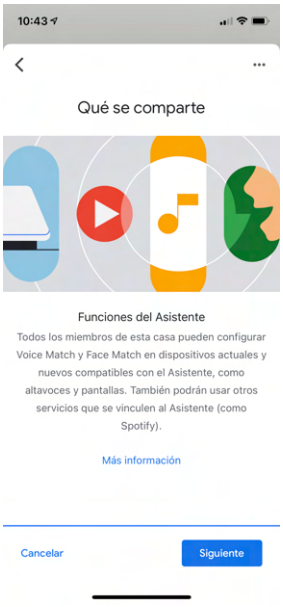


Figura 3.39: Funciones del asistente que se comparten con invitados



3.3.1.2. Instalación y edición de nuevos dispositivos conectados

Se permite añadir nuevos dispositivos y accesorios desde la aplicación *Google Home* para poder interactuar con ellos con el asistente. El proceso para añadir un nuevo dispositivo se describe en las Figuras 3.40, 3.41 y 3.42.

Figura 3.40: Acceso a opciones para añadir dispositivos

Figura 3.41: Opciones para añadir a la aplicación

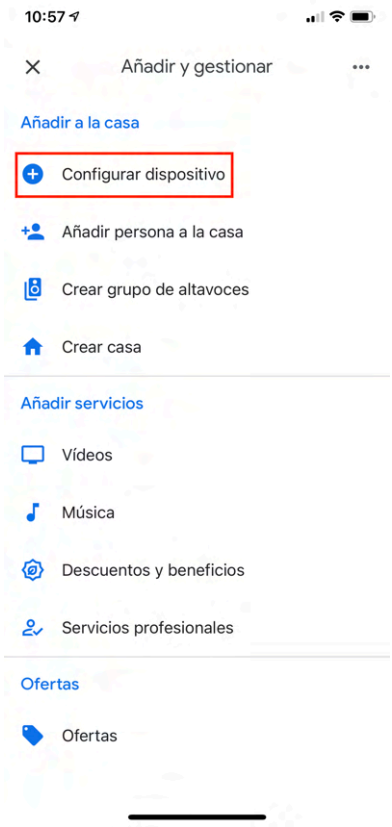


Figura 3.42: Configuración del nuevo dispositivo

No existe ningún método de control para determinar que usuarios de la casa tienen permiso para añadir o editar dispositivos. Por defecto, todos los usuarios que se inviten a la casa tienen el nivel de administradores y no existe opción para configurar los permisos a fecha actual, por lo que cualquier usuario podría hacer cambios en los dispositivos. En las Figuras 3.43, 3.44, 3.45, 3.46 y 3.47 se muestra el acceso a los ajustes de la unidad familiar donde se indican que usuarios forman parte de la casa y el nivel de permisos que tienen. Como se puede observar en la Figura 3.47, donde se resume el nivel de permisos en cuanto al acceso a dispositivos, cualquier usuario puede controlar, añadir y eliminar cualquier dispositivo de la casa.

Figura 3.43: Acceso a los ajustes de la casa

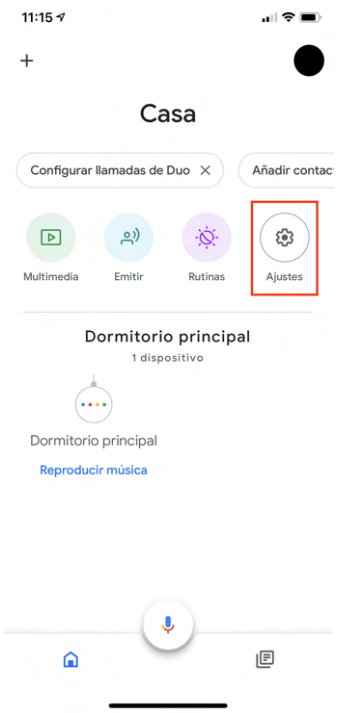


Figura 3.44: Acceso a la unidad familiar

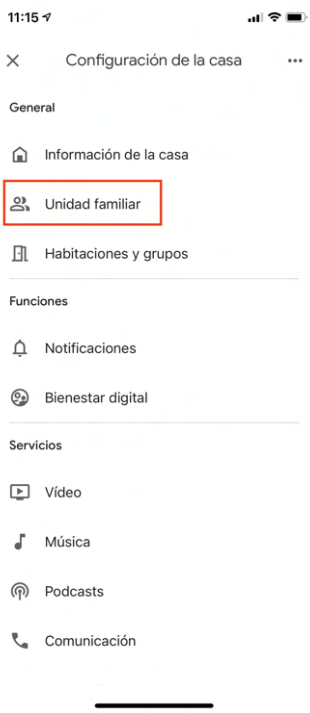


Figura 3.45: Vista de usuarios de la unidad familiar

Figura 3.46: Perfil de un usuario invitado

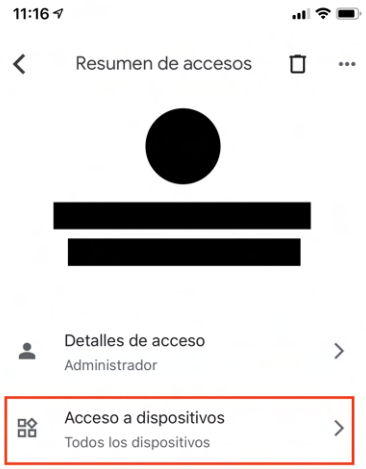


Figura 3.47: Detalle de permisos de acceso a dispositivos

3.3.1.3. Instalación de acciones de terceros

No existe un proceso de instalación al uso para añadir acciones (término con el que *Google* define las *skills*) de terceros al asistente, simplemente es necesario conocer el nombre con el que se invoca a la acción para utilizarla. Para visualizar las acciones de terceros en *iOS*, es necesario descargar la aplicación *Asistente de Google* [Figura 3.48].

Figura 3.48: Aplicación *Asistente de Google* en la *App Store*

En la aplicación, desde el menú *Explorar* se pueden buscar y guardar acciones en el asistente de *Google*, que quedarán vinculadas a la cuenta del usuario [Figuras 3.49, 3.50 y 3.51]. Si una acción necesita de ajustes adicionales, como iniciar sesión en un servicio, una vez esté configurada, todos los usuarios podrán hacer uso de la misma desde el *Google Home Mini*.

Figura 3.49: Pantalla de inicio de *Asistente de Google*

Figura 3.50: Exploración de acciones en el *Asistente de Google*

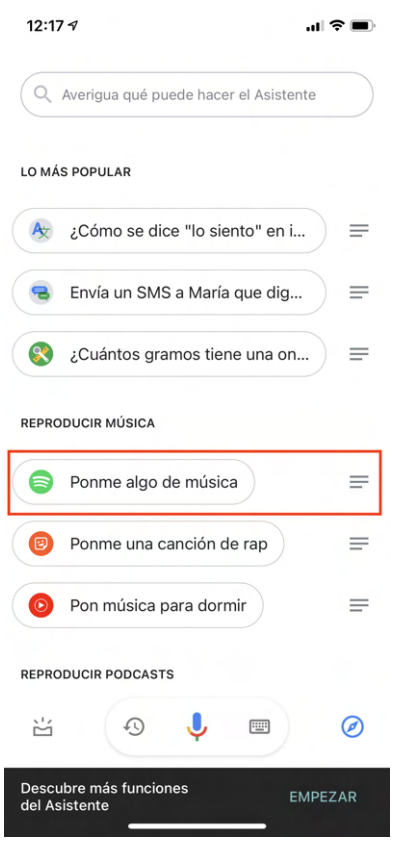
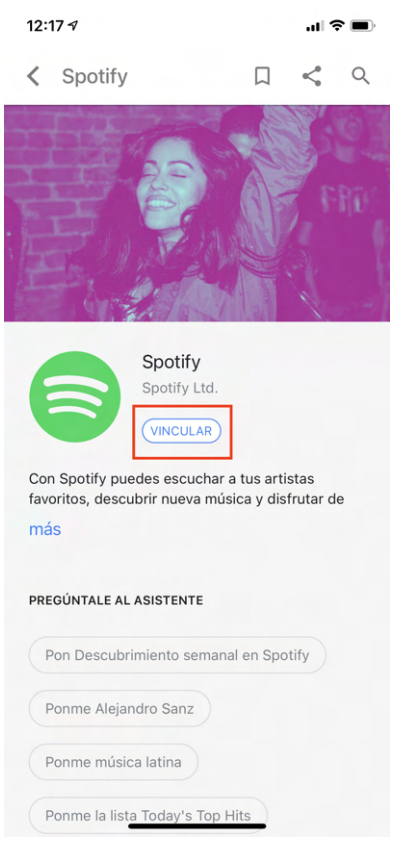


Figura 3.51: Vinculación de una acción de terceros al asistente



3.3.2. Pruebas de interacción

3.3.2.1. Interacción con el SPA y dispositivos conectados

Desde la aplicación no existen opciones para controlar el estado del *Google Home Mini*, pero sí se dispone de un botón físico para desconectar el micrófono del dispositivo. tal como se mostró en la introducción del dispositivo al comienzo de la Sección 3.3.

Como se mostró en la Figuras 3.46 y 3.47, no existen opciones de configuración de permisos de los usuarios que forman parte de la casa para controlar el uso que pueden hacer de los dispositivos. Los usuarios añadidos se convierte automáticamente en administradores con acceso completo tanto al SPA como a los accesorios conectados.

Para usuarios conectados a la misma red *Wi-Fi* que el *Google Home Mini* que no están registrados en la casa, se permite configurar en el dispositivo si se habilita el control de la reproducción de contenido multimedia en el *Google Home Mini* [Figuras 3.52, 3.53, 3.54 y 3.55].

Figura 3.52: Acceso al *Google Home Mini*

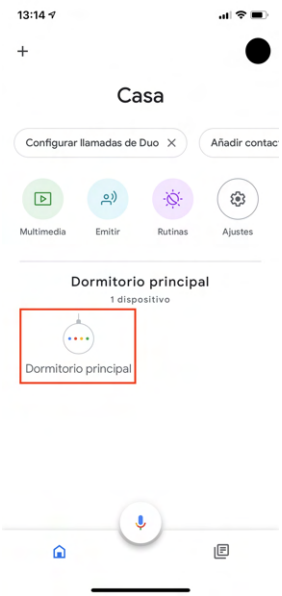


Figura 3.53: Acceso a los ajustes del *Google Home Mini*

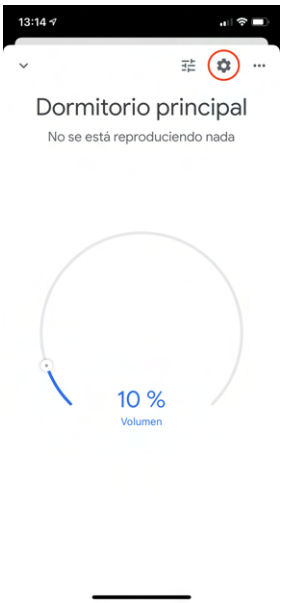


Figura 3.54: Acceso a los ajustes de reconocimiento y datos compartidos

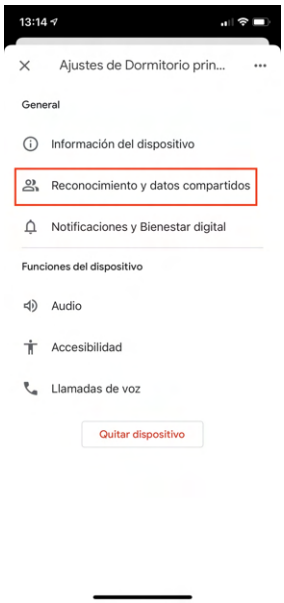
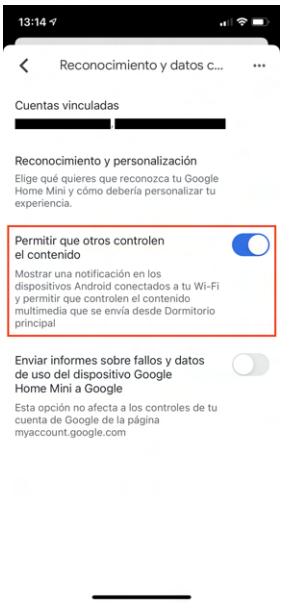


Figura 3.55: Opciones de control de reproducción de contenido



El único método para la aplicación de restricciones de uso es a través de la creación de una familia en la aplicación *Family Link* [Figura 3.56] de *Google* y el establecimiento de filtros de contenido en los ajustes del *Google Home Mini*, como se muestra en las Figuras 3.57, 3.58, 3.59, 3.60, 3.61, 3.62, 3.63 y 3.64.

Figura 3.56: Aplicación *Family Link* en la *App Store*

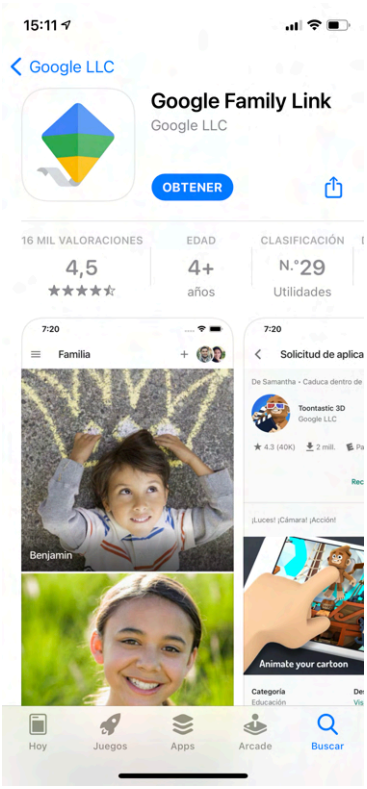


Figura 3.57: Acceso a notificaciones y bienestar digital

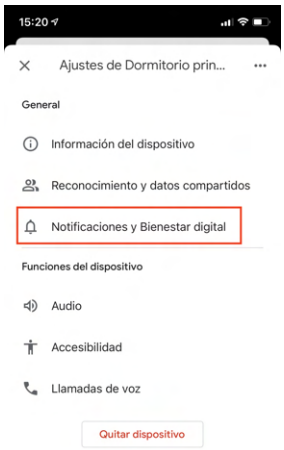


Figura 3.58: Acceso a bienestar digital

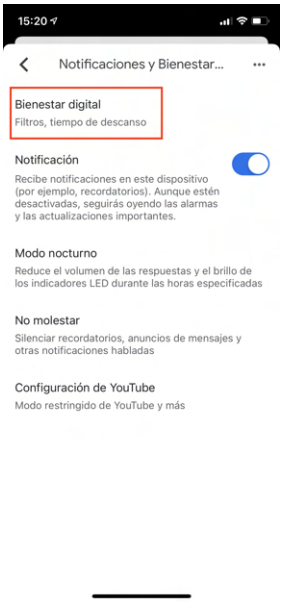


Figura 3.59: Información de la funcionalidad de bienestar digital

Figura 3.60: Información de la funcionalidad de filtros

Figura 3.61: Opciones de control de usuarios

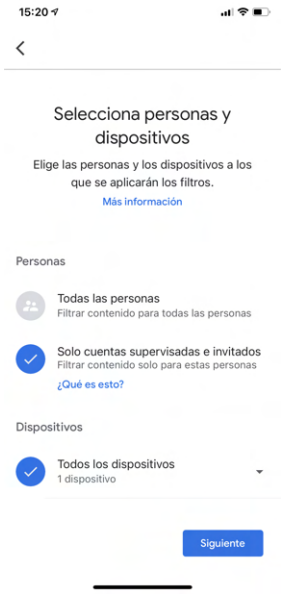


Figura 3.62: Opciones de control de vídeo

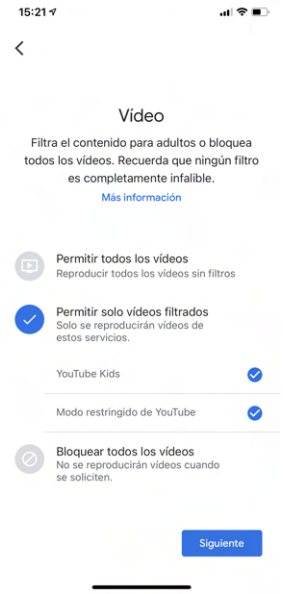


Figura 3.63: Opciones de control de música

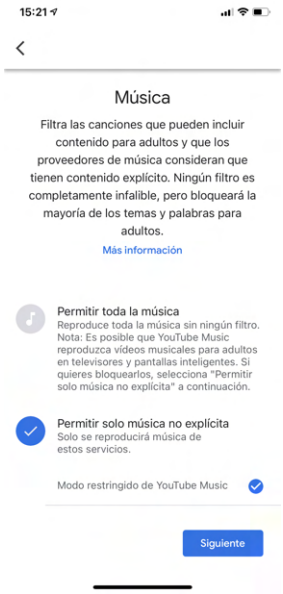
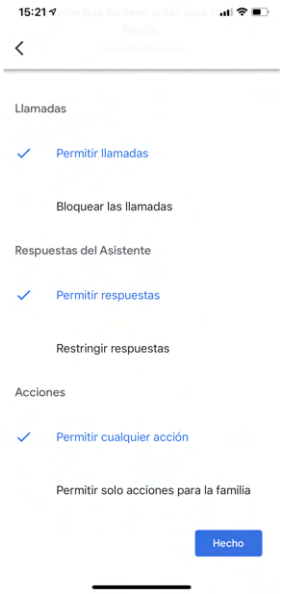


Figura 3.64: Controles adicionales



Tal como se puede observar en las Figuras 3.61, 3.62, 3.63 y 3.64, se puede configurar a que grupo de usuarios se aplicarán las restricciones, que contenido multimedia se permite y que comportamiento tendrá el asistente a la hora de realizar llamadas, proporcionar respuestas o ejecutar acciones. Los controles se aplican a todo el grupo de usuarios, no habiendo posibilidad de particularizarlos para personas específicas.

3.3.2.2. Interacción con acciones de terceros

Al no existir un proceso de instalación de acciones de terceros en el asistente de *Google*, no se ofrece ninguna opción para saber que aplicaciones están activas y definir que usuarios pueden hacer uso de las mismas desde la aplicación *Google Home*.

Como en el caso anterior, la única forma de restringir el uso de acciones es a través de la configuración de filtros en el *Google Home Mini*, tal como se muestra en la Figura 3.64. De nuevo, estas restricciones afectarán a todo el grupo de usuarios, no habiendo forma de individualizar los controles.

3.3.3. Pruebas de funcionalidad

3.3.3.1. Pagos y transacciones

Se puede configurar un método de pago para realizar transacciones con el asistente. La aplicación permite al usuario activar o desactivar los pagos en cualquier momento, así como configurar un método de autenticación adicional basado en el *hardware* del dispositivo móvil en el que la aplicación *Google Home* está instalada.

Figura 3.65: Acceso a ajustes del asistente

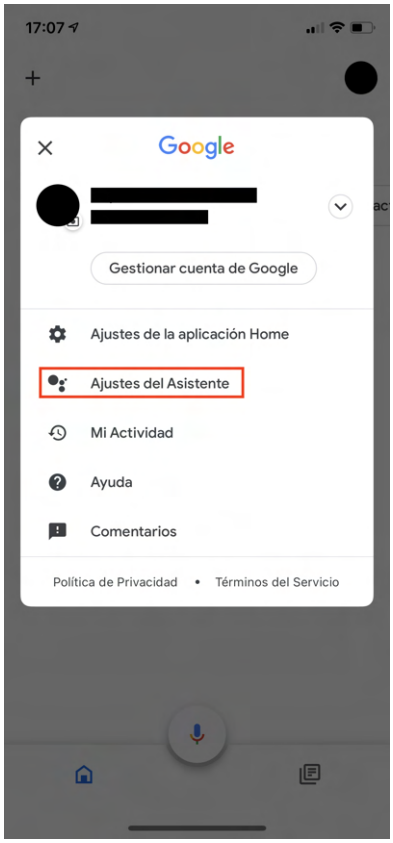


Figura 3.66: Acceso a ajustes de pagos

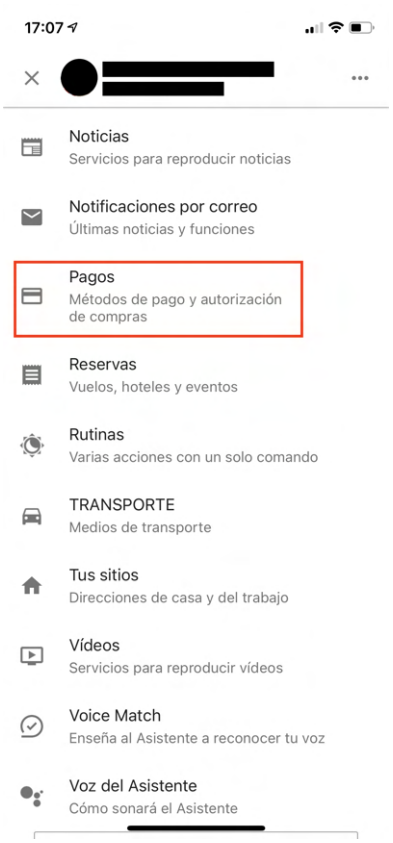
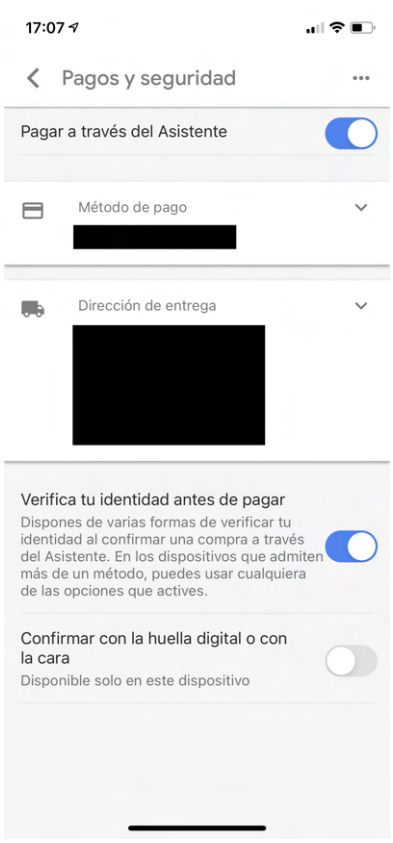


Figura 3.67: Configuración de pagos con el asistente



3.3.3.2. Posibilidad de crear perfiles "seguros" para menores

No se pueden crear perfiles para menores desde la aplicación *Google Home* para *iOS*. Esta funcionalidad sólo está disponible en dispositivos *Android* [44]. La única opción disponible para controlar el contenido al que los usuarios de la casa pueden acceder desde el asistente consiste de nuevo en crear filtros de contenido en los ajustes del asistente, tal como se describió en la Sección 3.3.2.1. Si bien esta opción restringe el contenido con el que se puede interactuar en el *Google Home Mini*, esto afecta a todos los usuarios a la vez, y otras opciones importantes quedan activadas, como la posibilidad de hacer pagos. Para restringir estas opciones, es necesario tener un dispositivo *Android* para dar de alta perfiles para menores en el *Google Home Mini* o desactivar también los pagos para todos los usuarios.

3.3.4. Pruebas de privacidad y seguridad

3.3.4.1. Control de respuestas que contengan datos personales

Desde los ajustes del *Google Home Mini* se pueden desactivar las respuestas que contengan datos personales, como eventos del calendario, mensajes, recordatorios, contactos, etc. El proceso se indica en las Figuras 3.68, 3.69, 3.70, 3.71 y 3.72.

Figura 3.68: Acceso al *Google Home Mini*

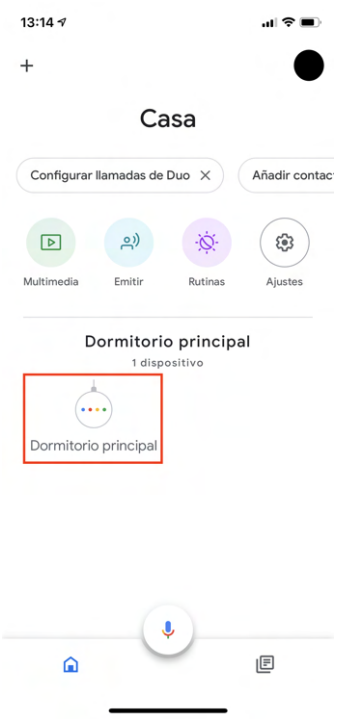


Figura 3.69: Acceso a los ajustes del *Google Home Mini*

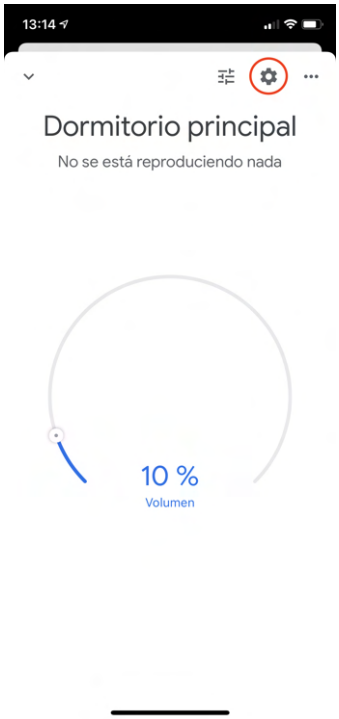


Figura 3.70: Acceso a ajustes de reconocimiento y datos

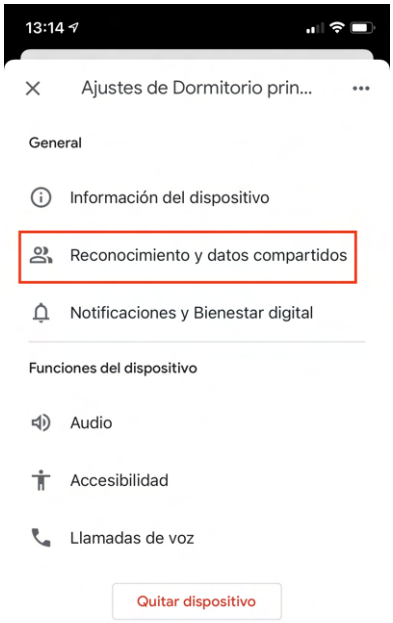


Figura 3.71: Acceso a reconocimiento y personalización

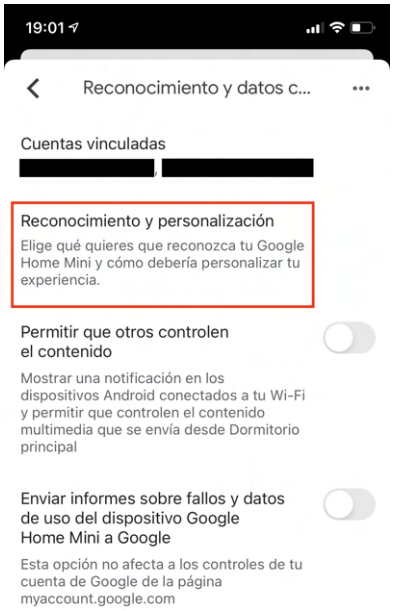
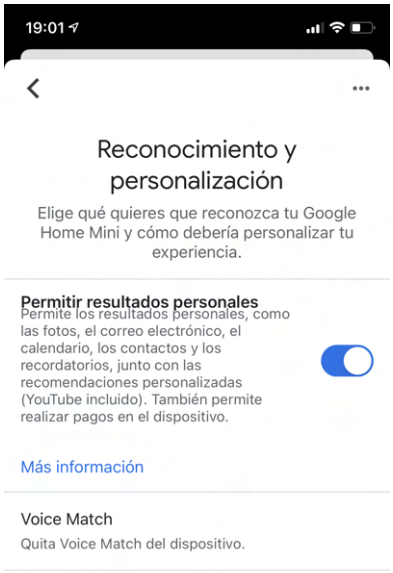


Figura 3.72: Control de respuestas personales

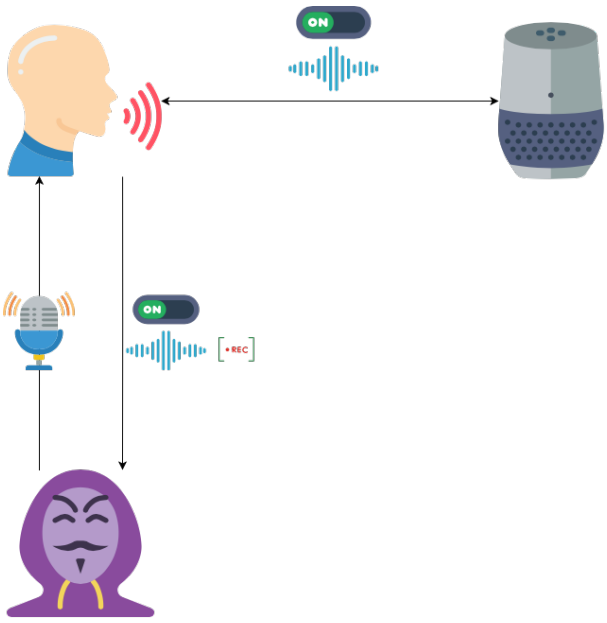


3.3.4.2. Métodos de autenticación

El *Google Home Mini* ofrece reconocimiento de voz, disponible en español, para distinguir a los usuarios que interactúen con el dispositivos. Se han realizado pruebas para comprobar la validez del reconocimiento de voz con dos voces en un rango de edad y timbre similar, siendo capaz el asistente de distinguir entre ambas.

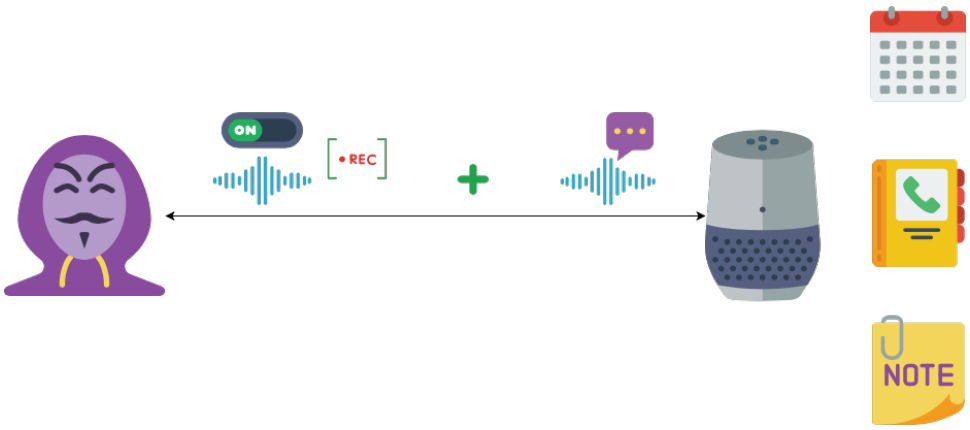
Tal como se advierte en la documentación del fabricante y en el aviso que se muestra cuando se configura *Voice Match* en el *Google Home Mini*, una grabación del usuario reproducida al dispositivo permite acceder a información personal, consiguiendo evadir la autenticación por voz. No es necesario tener el mensaje completo grabado con la voz del usuario legítimo, únicamente se necesita una grabación del usuario con el mensaje de activación `OK, Google` [Figura 3.73], pudiendo realizar después cualquier consulta, ya que no se verifica la voz del resto del mensaje, consiguiendo con éxito suplantar al usuario y acceder a su información [Figura 3.74].

Figura 3.73: Captura del mensaje de activación de un usuario legítimo



[1]

Figura 3.74: Ataque de suplantación con grabación del mensaje de activación



[1]

3.3.4.3. Filtrado de voz no humana (voz de usuario pregrabada y sistemas Text To Speech)

Como se ha expuesto en la sección anterior, un mensaje con la voz grabada de un usuario del *Google Home Mini* permite activarlo e interactuar con él. El dispositivo tampoco filtra mensajes procedentes de sistemas *Text To Speech*, como se ha comprobado reproduciendo un mensaje de este tipo al asistente con la palabra de activación y una consulta estándar.

[1] Iconos creados por *Smash icons* en flaticon.com

3.3.4.4. Posibilidad de interactuar con el historial de conversaciones con el asistente

Desde la aplicación *Google Home* se facilita acceso a la web *Mi Actividad* de *Google*, donde se puede visualizar, controlar y eliminar la actividad de la cuenta [Figuras 3.75 y 3.76]. Las opciones que ofrece *Google* para interactuar con el historial son más completas que las que *Apple* incluye en el *HomePod Mini*. Como se muestra en la Figura 3.76, se permite al usuario:

- Visualizar la actividad. Se puede observar la actividad con el asistente, pudiendo ver las conversaciones previas mantenidas.
- Pausar la actividad. Se permite desactivar el guardado de información sobre las interacciones que el usuario ha tenido con el asistente.
- Borrado automático. Existe una opción adicional para configurar un borrado automático de la actividad con el asistente, en plazos de 3, 18 y 36 meses.

Figura 3.75: Acceso a ajustes de *Mi Actividad* desde *Google Home*

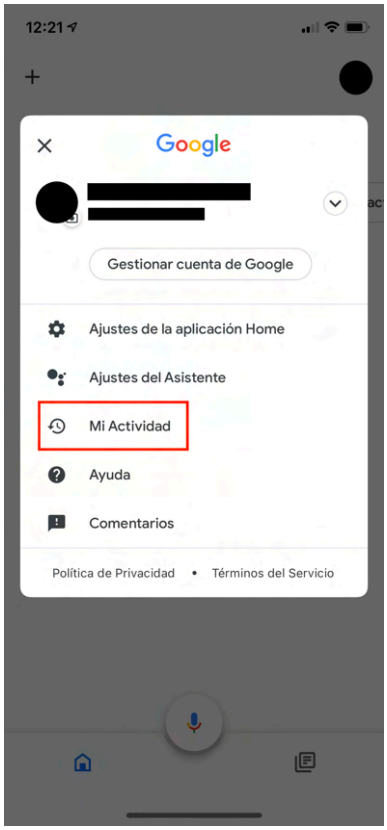


Figura 3.76: Opciones de control de información en *Mi Actividad*

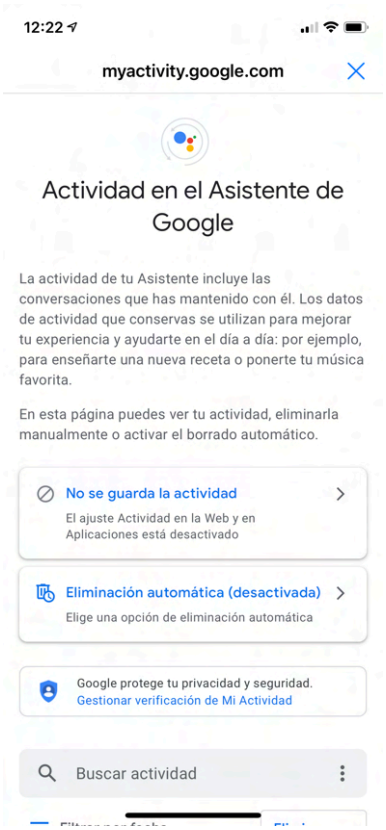


Tabla 3.3: Resumen de resultados de pruebas en el *Google Home Mini*

Característica	Resultado obtenido
Instalación	
Proceso de instalación del SPA	Es necesaria una cuenta de *Google*. Los datos de *Wi-Fi*, preferencias de la cuenta de *Google* y ajustes adicionales se comparten desde el dispositivo móvil
Instalación de nuevos dispositivos conectados	No se pueden configurar permisos para los usuarios de la casa. Cualquiera puede añadir o editar accesorios
Instalación de *skills* de terceros	No hay un proceso de instalación de acciones (*skills*) de terceros. Conociendo la frase de activación puede usarse cualquier acción
Interacción	
Interacción con el SPA y dispositivos conectados	Se puede desactivar la reproducción de contenido multimedia en el dispositivo. No se pueden definir permisos desde *Google Home*, siendo necesario crear una familia en la aplicación externa *Family Link* y configurar filtros por dispositivo en la aplicación *Google Home*
Interacción con *skills* de terceros	No hay controles de acciones de terceros por usuario, es necesario generar un filtro de contenido para todo el grupo de usuarios de la casa
Funcionalidad	
Pagos y transacciones	Se pueden configurar pagos desde el asistente. Soportan un método de autenticación adicional basado en el *hardware* del dispositivo donde se haya instalado la aplicación *Google Home*
Posibilidad de crear perfiles "seguros" para menores	Opción sólo disponible en *Android*. En el resto de dispositivos es necesario crear filtros de contenido de forma manual que afectan a todos los usuarios por igual, aunque opciones como pagos siguen estando activas
Privacidad y seguridad	
Respuestas que contengan información personal	Se permite desactivar las respuestas personales en los ajustes del *Google Home Mini*
Métodos de autenticación	El dispositivo permite autenticación por voz, aunque contando con una grabación de un usuario invocando al asistente se le puede suplantar, pudiendo realizar cualquier consulta
Filtrado de voz no humana (voz de usuario pregrabada, sistemas TTS)	No se filtran mensajes procedentes de grabaciones ni de voces sintéticas
Interacción con el historial de conversaciones	Se incluyen opciones completas para visualizar, pausar y eliminar el historial de forma automática

3.4. Amazon Echo Show 5 (2019)

El asistente personal de *Amazon* es el más usado, con una cuota de mercado estimada del 31.7 % en 2019 [16], 9500 marcas compatibles [10] y más de 77000 *skills* disponibles en 2020 [45]. Como *Google*, *Amazon* cuenta con su propio ecosistema de dispositivos que integran a *Alexa*, además de una variedad de equipos complementarios como enchufes inteligentes, cámaras, timbres o *routers* [46]. Todos los dispositivos que integran *Alexa* cuentan con micrófonos y algunos modelos también disponen de cámara (*Echo Show 10* [47], entre otros). *Alexa* también permite la integración de funcionalidades de terceros que se pueden invocar desde el asistente, denominadas *skills*.

El modelo analizado es el *Amazon Echo Show 5* de 1ª generación [Figuras 3.77 y 3.78], lanzado en Junio de 2019. El dispositivo cuenta con los controles habituales de volumen e indicadores del estado de funcionamiento. Es el único de los tres dispositivos analizados que cuenta con una pantalla táctil como método de presentación de información e interacción con el asistente, además de contar con cámara integrada. Dispone también de controles dedicados para desactivar los micrófonos del dispositivo y una tapa física para apagar la cámara.

Figura 3.77: *Echo Show 5*

Figura 3.78: Vista superior del *Echo Show 5*

Las pruebas se han realizado en las mismas condiciones que las anteriores, enlazando el dispositivo con un *iPhone X* con versión *iOS 14.6*.

En las secciones siguientes se incluye el detalle de las pruebas, junto con el resumen en la Tabla 3.4.

3.4.1. Pruebas de instalación

3.4.1.1. Instalación del SPA

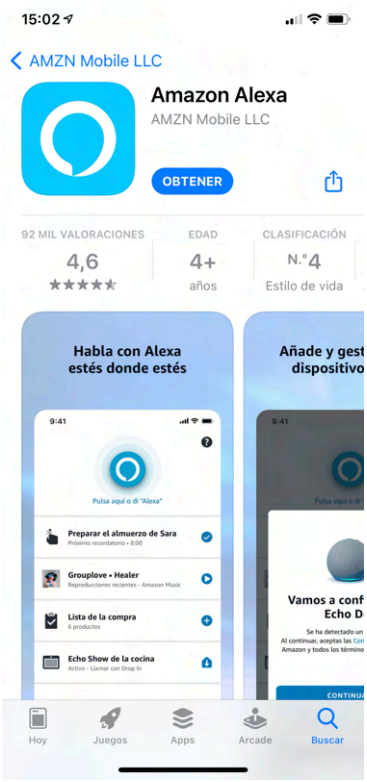
Para realizar los primeros ajustes en el dispositivo, no es necesario utilizar un dispositivo externo. Al disponer de pantalla, los datos de conexión a la red *Wi-Fi* se introducen directamente en el dispositivo. Se da la opción al usuario de almacenar la contraseña de la red en los servidores de *Amazon* para configurar futuros dispositivos de forma más rápida o continuar sin guardarla.

Una vez se conecta el dispositivo a la red, es obligatorio iniciar sesión con una cuenta de *Amazon* en el *Echo Show 5*. Cuando se inicia sesión en el dispositivo, queda vinculado a la cuenta de *Amazon* introducida.

Después se da la opción de guardar una ubicación en el dispositivo y personalizar la apariencia de la pantalla principal.

Para realizar cambios en los ajustes del asistente de *Amazon*, *Alexa*, es necesario descargar la aplicación del mismo nombre en un dispositivo móvil [Figura 3.79], cuya última versión disponible es la 2.2.422194.0.

Figura 3.79: Aplicación *Alexa* en la *App Store*



En la sección de privacidad de la aplicación, se incluye un informe con los datos que el desarrollador (*Amazon*) expone que utiliza cuando se hace uso de ella. De nuevo, se observa una recogida de datos más significativa que la realizada por la aplicación *Casa* de *Apple* y mayor que la de *Google Home*. La aplicación *Amazon Alexa* recoge los siguientes datos:

- Datos vinculados con el usuario.
  - Publicidad o marketing del desarrollador.
    - Historial de compras.
    - Datos de contacto, como dirección física, correo electrónico, nombre y número de teléfono.
    - Historial de búsqueda.
    - Identificadores del usuario y/o del dispositivo.
    - Datos de uso, como interacción con el producto y datos de publicidad
  - Análisis de datos.
    - Salud y forma física.
    - Historial de compras.
    - Ubicación exacta y/o aproximada.
    - Datos de contacto, como dirección física, correo electrónico, nombre, número de teléfono y otros datos.
    - Contactos.
    - Contenido del usuario, como correos electrónicos, mensajes de texto, fotos, vídeos, datos de audio, actividad en juegos, atención al cliente y otro contenido del usuario.
    - Historial de búsqueda.
    - Identificadores del usuario y/o del dispositivo.
    - Datos de uso, como interacción con el producto, datos de publicidad y otros datos.
    - Datos sensibles.
    - Diagnósticos, como datos de errores, datos de rendimiento y otros datos.
    - Otros tipos de datos.

- Personalización del producto.
  - Salud y forma física.
  - Historial de compras.
  - Ubicación exacta y/o aproximada.
  - Datos de contacto, como dirección física, correo electrónico, nombre, número de teléfono y otros datos.
  - Contactos.
  - Contenido del usuario, como correos electrónicos, mensajes de texto, fotos, vídeos, datos de audio, actividad en juegos, atención al cliente y otro contenido del usuario.
  - Historial de búsqueda.
  - Identificadores del usuario y/o del dispositivo.
  - Datos de uso, como interacción con el producto y otros datos,
  - Datos sensibles.
  - Otros tipos de datos
- Funcionalidad de la aplicación.
  - Salud y forma física.
  - Historial de compras.
  - Ubicación exacta y/o aproximada.
  - Datos de contacto, como dirección física, correo electrónico, nombre, número de teléfono y otros datos.
  - Contactos.
  - Contenido del usuario, como correos electrónicos, mensajes de texto, fotos, vídeos, datos de audio, actividad en juegos, atención al cliente y otro contenido del usuario.
  - Historial de búsqueda.
  - Identificadores del usuario y/o del dispositivo.
  - Datos de uso, como interacción con el producto, datos de publicidad y otros datos.
  - Datos sensibles.
  - Diagnósticos, como datos de errores, datos de rendimiento y otros datos.
  - Otros tipos de datos.
- Otros fines.
  - Salud y forma física.
  - Historial de compras.
  - Ubicación exacta y/o aproximada.
  - Datos de contacto, como dirección física, correo electrónico, nombre, número de teléfono y otros datos.
  - Contactos.
  - Contenido del usuario, como correos electrónicos, mensajes de texto, fotos, vídeos, datos de audio, actividad en juegos, atención al cliente y otro contenido del usuario.
  - Historial de búsqueda.
  - Identificadores del usuario y/o del dispositivo.
  - Datos de uso, como interacción con el producto, datos de publicidad y otros datos.
  - Datos sensibles.
  - Diagnósticos, como datos de errores, datos de rendimiento y otros datos.
  - Otros tipos de datos.

Una vez descargada la aplicación, es necesario proporcionar permisos de acceso al *Bluetooth* del *smartphone* e iniciar sesión con la cuenta de *Amazon* que se utilizó en el *Echo Show 5*.

La primera opción que se presenta en la aplicación es la de configurar un dispositivo nuevo. Al estar vinculado el *Echo Show 5* a la misma cuenta, se añade automáticamente a la aplicación [Figuras 3.80, 3.81 y 3.83].

Figura 3.80: Configuración de un nuevo dispositivo en *Alexa*

Figura 3.81: Búsqueda de un nuevo dispositivo en *Alexa*

Figura 3.82: Sección de dispositivos de la casa en *Alexa*

Figura 3.83: Información de dispositivos *Echo*

Desde la aplicación *Amazon Alexa* no se permite crear perfiles para usuarios diferentes para controlar los dispositivos. El usuario que ha registrado los dispositivos es el único que puede administrarlos, pero todas las personas que se encuentren en la casa o al alcance del *Echo Show 5* podrán activar *Alexa* y realizar peticiones [48].

Adicionalmente, el usuario puede crear un perfil de voz desde la aplicación, que permitirá personalizar la experiencia de uso en determinadas circunstancias, como la reproducción de contenido multimedia basado en los gustos del usuario que invocó el servicio, personalización de respuestas de *skills* o autenticación al realizar pagos por voz [Figuras 3.84, 3.85, 3.86 y 3.87].

Figura 3.84: Acceso a configuración

Figura 3.85: Acceso a configuración de cuentas

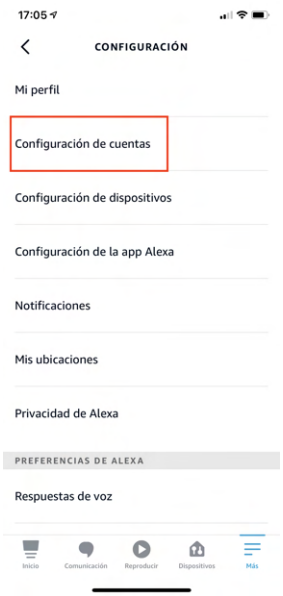


Figura 3.86: Acceso a voces reconocidas

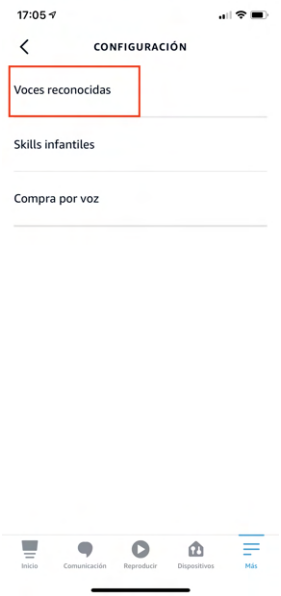


Figura 3.87: Información sobre el perfil de voz

Se pueden crear varios perfiles de voz para distinguir entre diferentes usuarios cuando interactúen con el *Echo Show 5*. Es un proceso más complejo comparado con la inserción de usuarios en el *HomePod Mini* o en el *Google Home Mini*, que requiere de otro dispositivo móvil con la aplicación *Amazon Alexa* instalada. En ella, el usuario administrador debe iniciar sesión y el nuevo usuario debe guardar un perfil de voz, que se sincronizará con el *Echo Show 5*.

3.4.1.2. Instalación y edición de nuevos dispositivos conectados

Se pueden añadir nuevos dispositivos y accesorios conectados al *Echo Show 5* desde la aplicación *Amazon Alexa* [Figuras 3.88, 3.89 y 3.90]. Como sólo el usuario administrador tiene acceso a la aplicación, es el único que puede añadir y editar dispositivos de forma externa. También se permite la detección e instalación de nuevos dispositivos a través del asistente, por lo que cualquier usuario que tenga al alcance el *Echo Show 5* podría añadir accesorios nuevos.

Figura 3.88: Acceso al menú para añadir dispositivos

Figura 3.89: Añadir un nuevo dispositivo

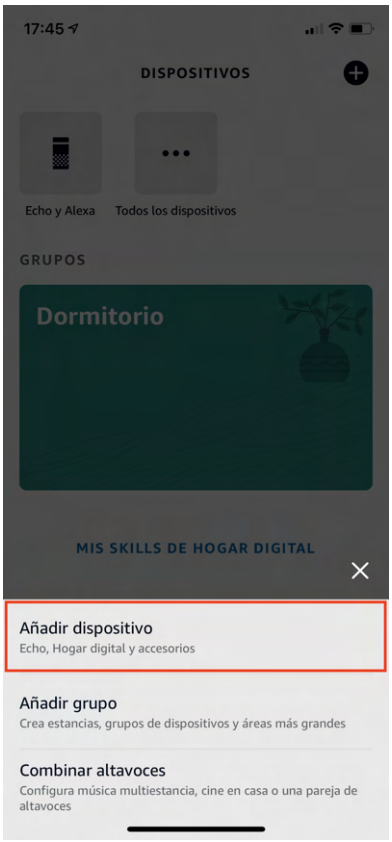
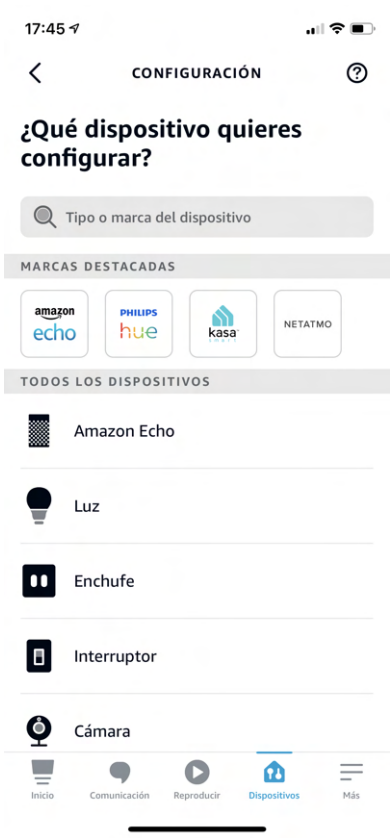


Figura 3.90: Selección del dispositivo nuevo



3.4.1.3. Instalación de skills de terceros

Las *skills* de terceros se instalan desde la aplicación móvil *Amazon Alexa*. Al contrario que en *Google Home*, es necesario dar de alta cada una de las *skills* que se vayan a usar antes de incorporarlas al asistente para invocarlas, lo que aporta una capa de seguridad adicional [Figuras 3.91, 3.92 y 3.93].

Figura 3.91: Acceso al *marketplace* de *skills*

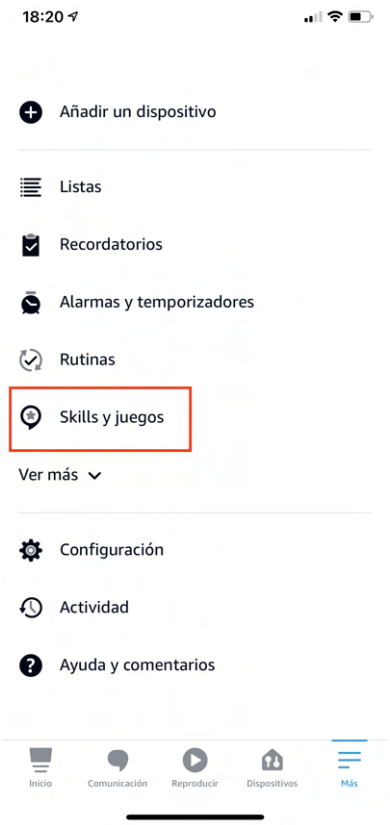
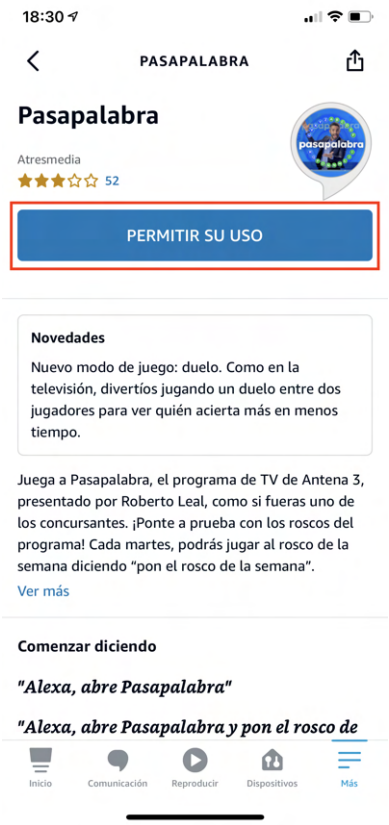


Figura 3.92: *Página principal del marketplace* de *skills*

Figura 3.93: Instalación de *skills* de terceros



Sin embargo, si se conoce la frase de activación de la *skill* es posible activarla interactuando directamente con *Alexa* en el *Echo Show 5*. Este proceso no requiere de ningún paso de verificación, por lo que cualquier usuario con acceso al asistente podría activar una *skill*.

3.4.2. Pruebas de interacción

3.4.2.1. Interacción con el SPA y dispositivos conectados

La interacción con el SPA se produce a través del control por voz, invocando al asistente con la palabra de activación, que puede cambiarse entre tres posibilidades: `Alexa`, `Amazon` o `Echo` [Figuras 3.94, 3.95, 3.96 y 3.97].

Como se ha descrito en la sección de introducción del dispositivo, el *Echo Show 5* cuenta con controles dedicados para apagar cámara y micrófono.

No hay posibilidad de distinguir comportamientos para diferentes usuarios al no haber una opción de registro de personas en la aplicación ni en el *Echo Show 5*.

Figura 3.94: Acceso a los dispositivos conectados

Figura 3.95: Acceso al *Echo Show 5*

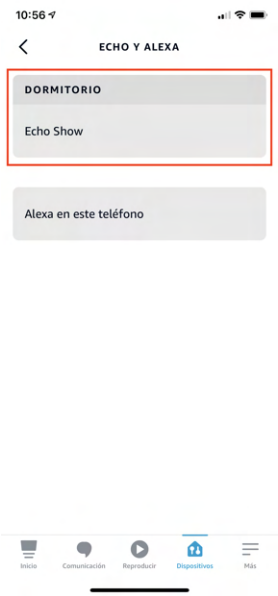


Figura 3.96: Acceso a los ajustes de activación

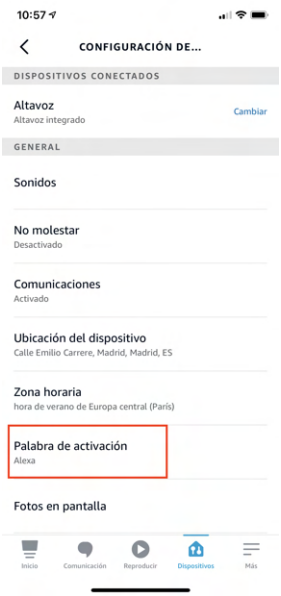
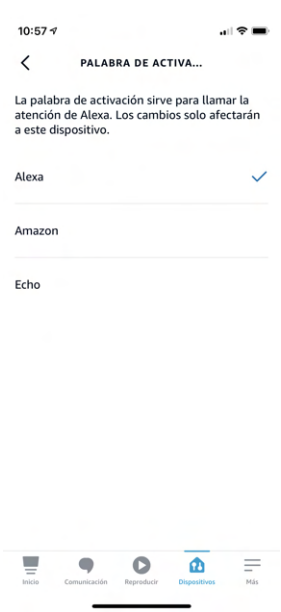


Figura 3.97: Opciones de configuración de la palabra de activación



3.4.2.2. Interacción con skills de terceros

Desde *Amazon Alexa* es posible configurar los datos que se comparten con cada *skill*, ver cuales están activadas y desactivarlas en cualquier momento. Sin embargo, al no haber registro de usuarios, no se puede diferenciar el uso de *skills* por usuario y, como se ha mostrado en la Sección 3.4.1.3, cualquier usuario puede activar de nuevo una *skill* deshabilitada interactuando con *Alexa* en el *Echo Show 5*.

3.4.3. Pruebas de funcionalidad

3.4.3.1. Pagos y transacciones

Se pueden realizar pagos con el asistente, que se activarán cuando se realicen pedidos en la tienda de *Amazon* o se utilicen ciertas *skills* que tenga funcionalidad de pagos, entre otras situaciones. La opción viene activada por defecto sin confirmación de compra, pudiendo configurarse desde la aplicación *Amazon Alexa* [Figuras 3.98, 3.99, 3.100 y 3.101]. Para utilizar los pagos desde el asistente, es necesario tener configurado el servicio *1 Click* o *Buy Now* de *Amazon* [49].

Figura 3.98: Acceso a los ajustes

Figura 3.99: Acceso a los ajustes de la cuenta

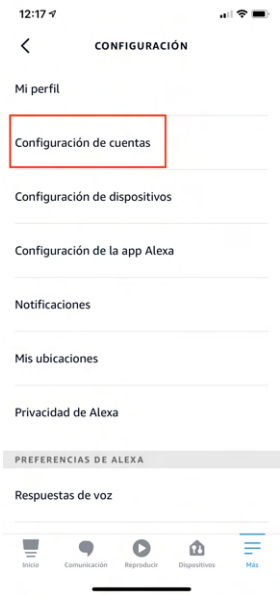


Figura 3.100: Acceso a compra por voz

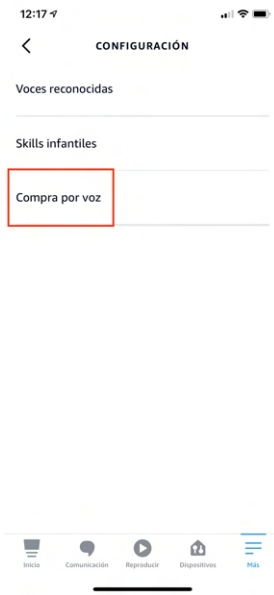
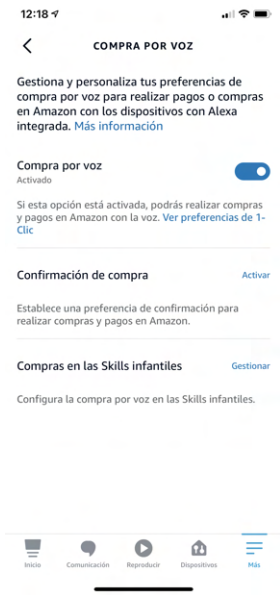


Figura 3.101: Configuración de la compra por voz



Se proporcionan varias opciones para hacer más seguro el proceso de compra en la sección de confirmación de la compra [Figuras 3.102 y 3.103]:

- `Perfil de voz`. Se utiliza el perfil de voz guardado para autenticar al usuario de forma automática a la hora de realizar el pago.
- `Código de voz`. Es necesario conocer un código de cuatro dígitos para realizar compras.

Figura 3.102: Acceso a opciones de confirmación de compra

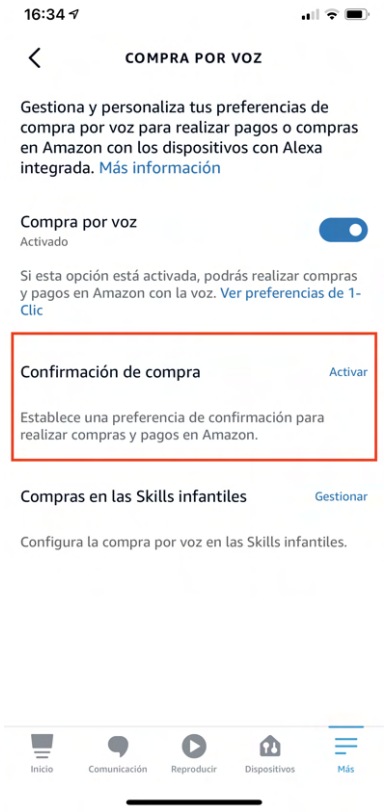
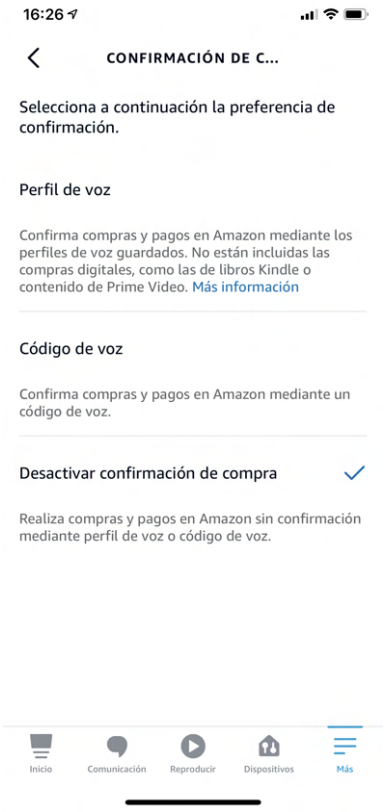


Figura 3.103: Configuración de confirmación de compra



3.4.3.2. Posibilidad de crear perfiles "seguros" para menores

No existe una opción dedicada para crear un perfil de uso para menores donde ciertas funcionalidades estén restringidas. Para controlar que se haga un uso seguro para menores, es necesario configurar las opciones de control de contenido del *Echo Show 5* teniendo que revisar varios menús.

Desde el punto de vista del contenido que se puede consumir o al que se puede acceder con el asistente, desde el propio *Echo Show 5* se puede restringir el uso de ciertas funcionalidades. En el menú *Configuración* > *Acceso restringido* se puede bloquear la visualización de tráileres de películas, el uso del navegador web incorporado, la posibilidad de buscar vídeos en Internet y el uso de ciertos proveedores de vídeo. Para realizar cambios en el apartado de *Acceso restringido* del *Echo Show 5* es necesario introducir la contraseña de la cuenta de *Amazon* asociada, por lo que un menor no podrá revertir los bloqueos impuestos.

Con respecto a los dispositivos que se pueden invocar desde el asistente, no hay ninguna forma de restringir el uso para los menores.

La misma situación se produce con las *skills* de terceros. No se puede controlar que *skills* puede invocar un menor. Además, si se desactiva una *skill* desde la aplicación *Amazon Alexa* o directamente desde el asistente, se puede volver a activar sin ningún tipo de validación, simplemente se confirmará que se quiere volver a activar la *skill* y esta podrá ser invocada de nuevo. Si existe un control sobre la categoría de *Skills infantiles*, aquellas que el desarrollador ha indicado como adecuadas para dicho colectivo. Desde la aplicación *Amazon Alexa* se puede activar o desactivar este tipo de *skills* [Figuras 3.104, 3.105, 3.106 y 3.107]. Además, también se da la opción de desactivar los pagos en *skills* para menores para aumentar la seguridad en el uso [Figuras 3.108, 3.109, 3.110, 3.111 y 3.112].

Figura 3.104: Acceso a los ajustes

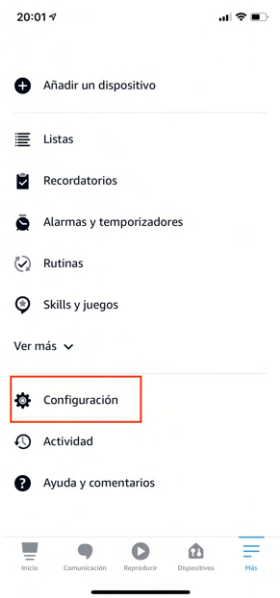


Figura 3.105: Acceso a los ajustes de la cuenta

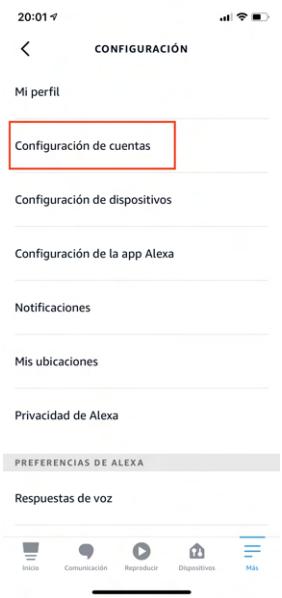


Figura 3.106: Acceso a *skills* infantiles

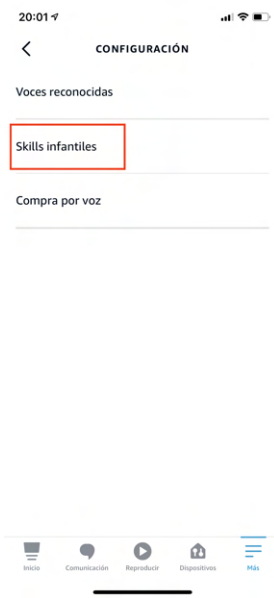


Figura 3.107: Configuración de *skills* infantiles

Figura 3.108: Acceso a los ajustes

Figura 3.109: Acceso a los ajustes de la cuenta

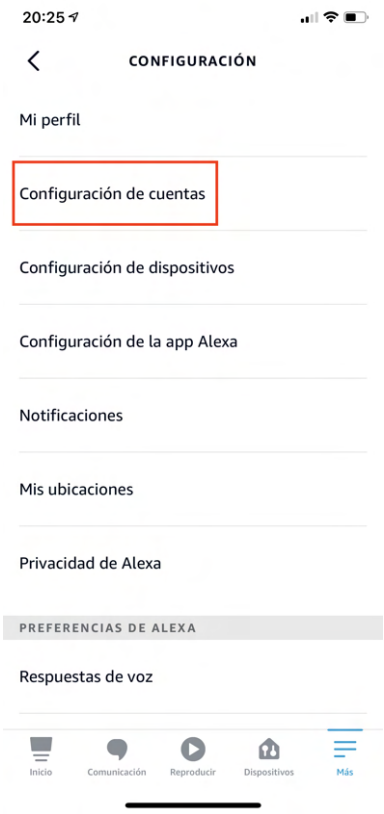


Figura 3.110: Acceso a ajustes de compras por voz

Figura 3.111: Acceso ajustes de compras en *skills* infantiles

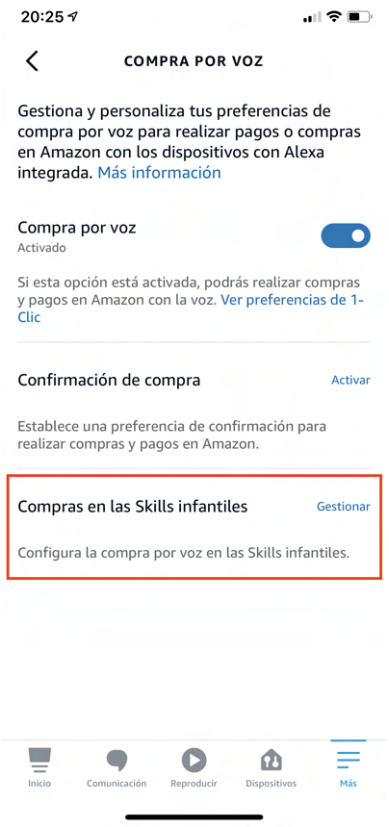
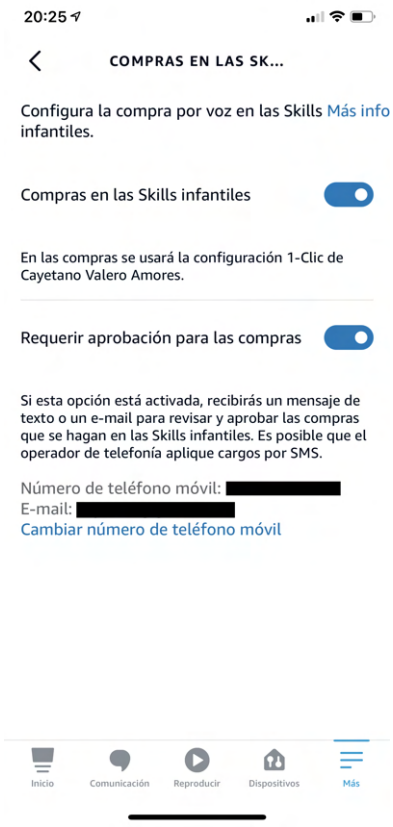


Figura 3.112: Ajustes de compras en *skills* infantiles



3.4.4. Pruebas de privacidad y seguridad

3.4.4.1. Control de respuestas que contengan información personal

No hay una opción dedicada para controlar las respuestas del asistente que puedan contener información personal como en *Siri* o *Google Assistant*. Para controlar que no se de cierta información cuando alguien interactúe con el asistente, la única opción es desactivar el servicio que puede contener esa información. Por ejemplo, si un usuario no quiere que alguien pueda acceder a sus eventos del calendario, debe desvincular su cuenta de calendario del *Echo Show 5* y desactivar la funcionalidad.

3.4.4.2. Métodos de autenticación

El *Echo Show 5* cuenta con reconocimiento de voz, pero sólo se utiliza para funciones de personalización, por ejemplo en *skills*. No se utiliza para tareas de seguridad, pudiendo cualquier persona acceder, por ejemplo, a los recordatorios o al calendario del usuario principal que configuró el dispositivo.

Además, el *Echo Show 5* es susceptible al mismo problema de autenticación descrito en el *Google Home Mini* [Figuras 3.73 y 3.74 en la Sección 3.3.4.2]. Teniendo una grabación de un usuario del *Echo Show 5* con la palabra de activación que tenga configurada, es posible suplantar al usuario reproduciendo la grabación, y a continuación enviando el comando deseado. Sólo se realiza reconocimiento de voz en la palabra de activación, por lo que una vez reconoce al usuario legítimo con la grabación, se puede enviar cualquier mensaje al asistente.

3.4.4.3. Filtrado de voz no humana (voz pregrabada, sistemas TTS)

Tal como se ha expuesto en el apartado anterior, el *Echo Show 5* no implementa un método de filtrado de mensajes grabados. Como en el resto de dispositivos analizados, también es posible activar el altavoz con una voz sintética que lea la palabra de activación y un mensaje cualquiera.

3.4.4.4. Posibilidad de interactuar con el historial de conversaciones con el asistente

Se ofrece un rango de opciones muy completo para controlar los datos del historial de conversaciones y la recolección de datos.

Desde el propio asistente se puede permitir que a través de una consulta se elimine el mensaje que se acaba de enviar, o los correspondientes a las 24 horas anteriores.

Si se necesita un control más exhaustivo de las opciones de historial, es necesario utilizar la aplicación *Amazon Alexa* [Figuras 3.113, 3.114 y 3.115].

Figura 3.113: Acceso a los ajustes generales

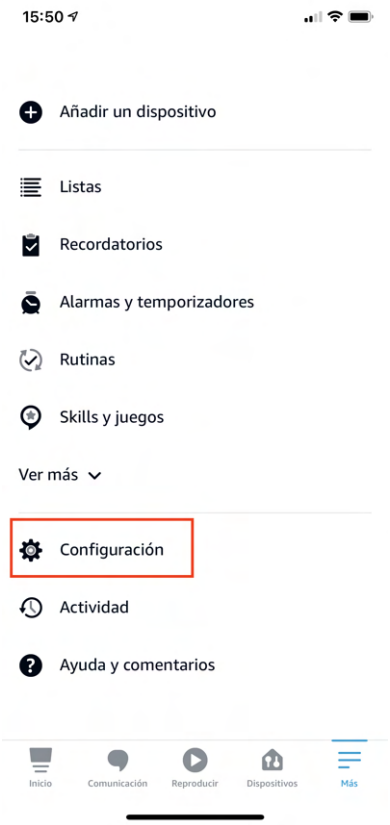


Figura 3.114: Acceso ajustes de privacidad de *Alexa*

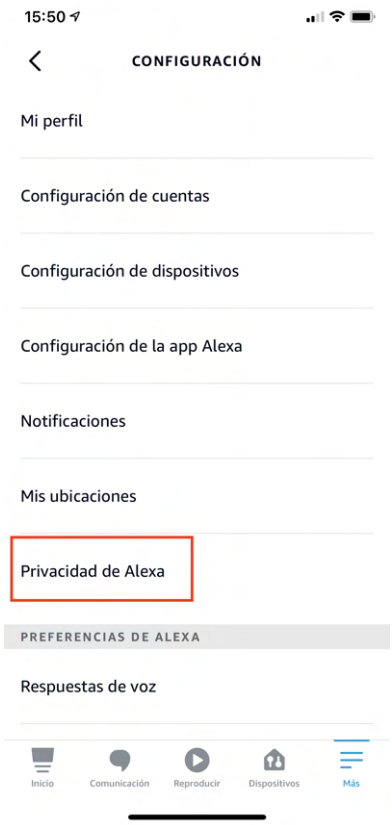
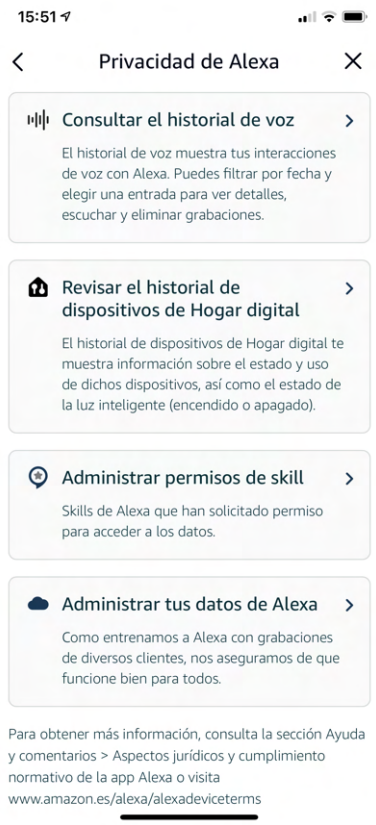


Figura 3.115: Configuración de privacidad



Como se puede observar en la Figura 3.115, se ofrece un abanico de opciones completo para visualizar el historial de conversaciones con el asistente en la opción `Consultar el historial de voz` [Figuras 3.116 y 3.117] y para administrar la recolección de datos y establecer un borrado automático de las conversaciones en el apartado `Administrar tus datos de Alexa` [Figuras 3.118 y 3.119].

Desde el apartado `Consultar el historial de voz` se pueden comprobar las interacciones con el asistente, tal como se muestra en la Figura 3.117. Además se proporciona una opción rápida para borrar todo el historial de las últimas 24 horas.

En la sección `Administrar tus datos de Alexa` [Figura 3.119] se establecen las directrices de almacenamiento y uso del historial de conversaciones con *Alexa*. En cuanto al almacenamiento, se permite al usuario establecer períodos de borrado automático de 3 y 18 meses, además de las opciones de mantener el historial de forma indefinida y la posibilidad de desactivar el historial de conversaciones.

Figura 3.116: Acceso al historial de conversaciones

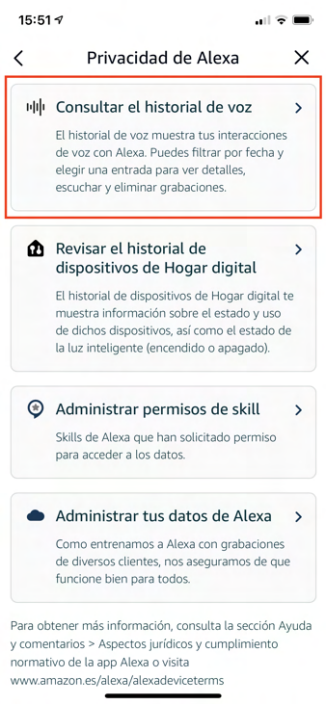


Figura 3.117: Historial de conversaciones con *Alexa*

Figura 3.118: Acceso a la gestión de los datos

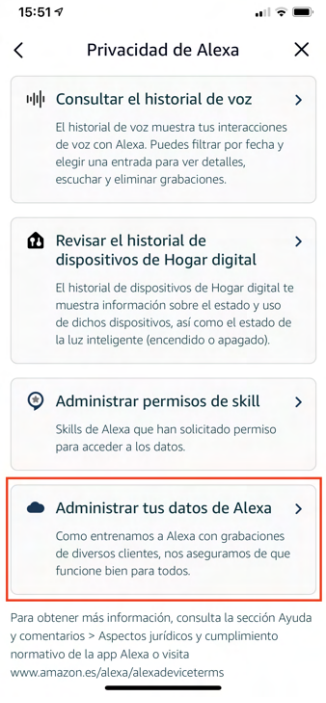


Figura 3.119: Administración de datos de *Alexa*

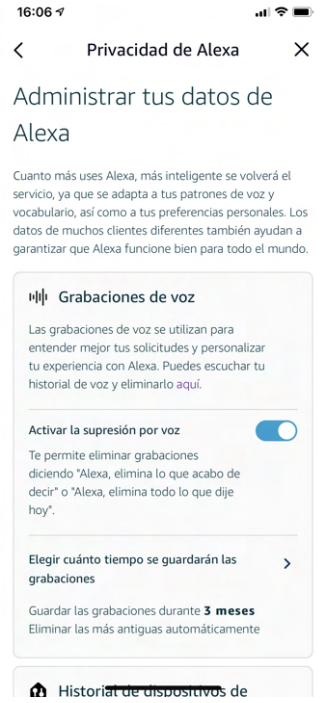


Tabla 3.4: Resumen de resultados de pruebas en el *Amazon Echo Show 5*

Característica	Resultado obtenido
Instalación	
Proceso de instalación del SPA	Es necesaria una cuenta de *Amazon* para utilizar el *Echo Show 5*. Los datos de *Wi-Fi* se introducen en el dispositivo y los ajustes de *Alexa* se sincronizan desde el dispositivo móvil
Instalación de nuevos dispositivos conectados	Sólo el administrador puede añadir dispositivos y editarlos con la aplicación *Alexa*
Instalación de *skills* de terceros	El administrador puede permitir el uso de *skills* desde la aplicación *Amazon Alexa* antes de usarlas, pero un usuario cualquiera puede activar una *skill* desde el *Echo Show 5* sin confirmación, incluso puede activar de nuevo una *skill* deshabilitada
Interacción	
Interacción con el SPA y dispositivos conectados	Se permite cambiar la palabra de activación entre tres posibilidades. El *Echo Show 5* dispone de controles para apagar cámara y micrófono. No hay posibilidad de distinguir entre usuarios ni de definir permisos para interacciones específicas
Interacción con *skills* de terceros	Desde la aplicación *Amazon Alexa* se puede configurar que datos se comparten con cada *skill* y ver cuáles están activadas. No se puede diferenciar el acceso a *skills* por usuario
Funcionalidad	
Pagos y transacciones	Se pueden realizar pagos con el asistente a través de *Amazon 1 Click*. Se pueden configurar métodos de confirmación adicionales a través del perfil de voz o de un código de cuatro dígitos
Posibilidad de crear perfiles "seguros" para menores	No hay una opción dedicada para establecer un perfil de uso seguro para menores. Es necesario desactivar y restringir ajustes en cada una de las categorías (reproducción de contenido multimedia, navegador web, pagos, *skills*, etc. No es posible restringir el uso de accesorios conectados a menores
Privacidad y seguridad	
Respuestas que contengan información personal	No existe la posibilidad de desactivar respuestas que contengan información personal. Es necesario desactivar las funcionalidades por completo, ya que cualquier usuario puede invocarlas
Métodos de autenticación	El *Echo Show 5* cuenta con reconocimiento de voz, pero sólo se utiliza para funciones de personalización, por ejemplo en *skills*. No se utiliza para tareas de seguridad
Filtrado de voz no humana (voz de usuario pregrabada, sistemas TTS)	No se filtran mensajes grabados ni aquellos procedentes de voces sintéticas
Interacción con el historial de conversaciones	Se permite el borrado de conversaciones desde el asistente y la aplicación. Ofrece la opción de borrado automático de las conversaciones y desactivar el guardado de grabaciones

Tabla 3.5: Comparativa de características de instalación e interacción de SPA comerciales

Característica	*Google Home Mini*	*Amazon Echo Show 5*	*Apple HomePod Mini*
Instalación			
Proceso de instalación del SPA	Es necesaria una cuenta de *Google*. Los datos de *Wi-Fi*, preferencias de la cuenta de *Google* y ajustes adicionales se comparten desde el dispositivo móvil	Es necesaria una cuenta de *Amazon* para utilizar el *Echo Show 5*. Los datos de *Wi-Fi* se introducen en el dispositivo y los ajustes de *Alexa* se sincronizan desde el dispositivo móvil	Es necesaria una cuenta de *iCloud* y un dispositivo de *Apple*. Los datos de *Wi-Fi*, *Siri* y preferencias se comparten desde el *iPhone*
Instalación de nuevos dispositivos conectados	No se pueden configurar permisos para los usuarios de la casa. Cualquiera puede añadir o editar accesorios	Sólo el administrador puede añadir dispositivos y editarlos con la aplicación *Alexa*	Se permite configurar si un usuario tiene permiso para añadir o editar dispositivos
Instalación de *skills* de terceros	No hay un proceso de instalación de acciones (*skills*) de terceros. Conociendo la frase de activación puede usarse cualquier acción	El administrador puede permitir el uso de *skills* desde la aplicación *Amazon Alexa* antes de usarlas, pero un usuario cualquiera puede activar una *skill* desde el *Echo Show 5* sin confirmación, incluso puede activar de nuevo una *skill* deshabilitada	No se pueden configurar *skills* de terceros
Interacción			
Interacción con el SPA y dispositivos conectados	Se puede desactivar la reproducción de contenido multimedia en el dispositivo. No se pueden definir permisos desde *Google Home*, siendo necesario crear una familia en la aplicación externa *Family Link* y configurar filtros por dispositivo en la aplicación *Google Home*	El *Echo Show 5* dispone de controles para apagar cámara y micrófono. No hay posibilidad de distinguir entre usuarios ni de definir permisos para interacciones específicas	Se puede desactivar la invocación por voz o el control táctil. Se permite desactivar el uso de dispositivos multimedia y particularizar por usuario el control de accesorios conectados
Interacción con *skills* de terceros	No hay controles de acciones de terceros por usuario, es necesario generar un filtro de contenido para todo el grupo de usuarios de la casa	El administrador puede permitir el uso de *skills* desde la aplicación *Amazon Alexa* antes de usarlas, pero un usuario cualquiera puede activar una *skill* desde el *Echo Show 5* sin confirmación, incluso puede activar de nuevo una *skill* deshabilitada	No hay posibilidad de interactuar con *skills* de terceros

Tabla 3.6: Comparativa de características de funcionalidad y privacidad y seguridad de SPA comerciales

Característica	*Google Home Mini*	*Amazon Echo Show 5*	*Apple HomePod Mini*
Funcionalidad			
Pagos y transacciones	Se pueden configurar pagos desde el asistente. Soportan un método de autenticación adicional basado en el *hardware* del dispositivo donde se haya instalado la aplicación *Google Home*	Se pueden realizar pagos con el asistente a través de *Amazon 1 Click*. Se pueden configurar métodos de confirmación adicionales a través del perfil de voz o de un código de cuatro dígitos	No se permiten los pagos o compras desde el *HomePod Mini*
Posibilidad de crear perfiles "seguros" para menores	Opción sólo disponible en *Android*. En el resto de dispositivos es necesario crear filtros de contenido de forma manual que afectan a todos los usuarios por igual, aunque opciones como pagos siguen estando activas	No hay una opción dedicada para establecer un perfil de uso seguro para menores. Es necesario desactivar y restringir ajustes en cada una de las categorías (reproducción de contenido multimedia, navegador web, pagos, *skills*, etc. No es posible restringir el uso de accesorios conectados a menores	No hay posibilidad de crear usuarios para menores, es necesario acceder a cada control y desactivarlo de forma manual
Privacidad y seguridad			
Respuestas que contengan información personal	Se permite desactivar las respuestas personales en los ajustes del *Google Home Mini*	No existe la posibilidad de desactivar respuestas que contengan información personal. Es necesario desactivar las funcionalidades por completo, ya que cualquier usuario puede invocarlas	Es necesario activar el reconocimiento de voz para dar respuestas con datos personales. No disponible en español
Métodos de autenticación	El dispositivo permite autenticación por voz, aunque contando con una grabación de un usuario invocando al asistente se le puede suplantar, pudiendo realizar cualquier consulta	El *Echo Show 5* cuenta con reconocimiento de voz, pero sólo se utiliza para funciones de personalización, por ejemplo en *skills*. No se utiliza para tareas de seguridad	Autenticación por reconocimiento de voz (no disponible en español)
Filtrado de voz no humana (voz de usuario pregrabada, sistemas TTS)	No se filtran mensajes procedentes de grabaciones ni de voces sintéticas	No se filtran mensajes grabados ni aquellos procedentes de voces sintéticas	Un mensaje de activación pregrabado o leído con un sistema TTS puede invocar al asistente
Interacción con el historial de conversaciones	Se incluyen opciones completas para visualizar, pausar y eliminar el historial de forma automática	Se pueden eliminar conversaciones recientes desde el asistente. En la aplicación se ofrecen opciones completas de visualización, borrado y pausado del historial de conversaciones, disponiendo de opciones de borrado automático	Se permite enviar una petición para eliminar el historial de los servidores. No se puede visualizar el historial de conversaciones

Capítulo 4

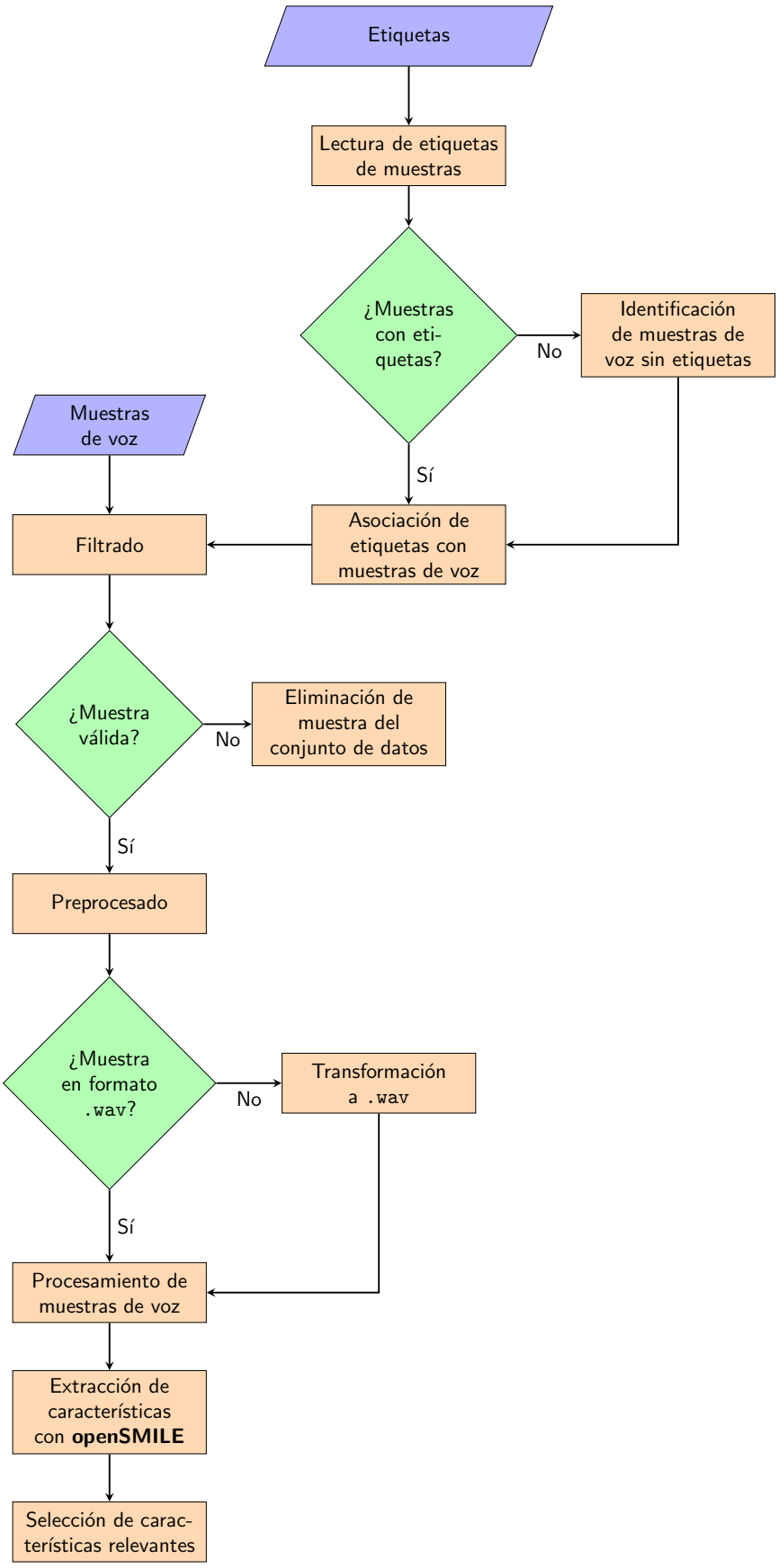
Desarrollo de herramienta de clasificación de voz para autenticación en SPA

El desarrollo de la herramienta de clasificación de voz se ha dividido en dos secciones. En la Sección 4.1 se recoge el detalle del proceso de preparación, filtrado y procesado de los conjuntos de datos. En la Sección 4.2 se describe el proceso de creación, entrenamiento y validación de los modelos de clasificación basados en Inteligencia Artificial.

4.1. Preprocesado y extracción de características

El primer paso para desarrollar el modelo de clasificación de voz para autenticación se centra en la preparación de los datos de los conjuntos *Common Voice* de *Mozilla* [Sección 2.3.3.1], para adecuarlos a las características requeridas por los algoritmos de Inteligencia Artificial que se usarán después. En la Figura 4.1 se muestra un esquema con los pasos que se han seguido para llevar a cabo el procesamiento de los conjuntos de datos, que se describe en detalle en los apartados siguientes. El proceso que se muestra en esta sección con el detalle del código corresponde al conjunto de datos *Common Voice* en inglés, siendo el preprocesado y extracción de características similar en el conjunto *Common Voice* en español.

Figura 4.1: Diagrama general de preparación y procesamiento del conjunto de datos



4.1.1. Carga de librerías

Para trabajar con el conjunto de datos *Common Voice* de *Mozilla* se ha hecho uso de diversas herramientas, que aportan las siguientes funcionalidades:

- Conversión de los archivos de audio a *.wav*. Para esta tarea se ha utilizado la librería *pydub*.
- Extracción de características de las muestras de voz. Tal como se mostró en la Sección 2.4.1, se ha hecho uso de la librería *openSMILE*.
- Trabajo con estructuras de datos basadas en *DataFrames*, para lo que se ha utilizado la librería *pandas*.

```
import time
import sox
import os

import matplotlib.pyplot as plt
import numpy as np
import pandas as pd

import audiofile
import opensmile

# Herramienta para conversión de .mp3 a .wav
import pydub
from pydub import AudioSegment

# Progress bar en Python https://github.com/niltonvolpato/python-progressbar
from progressbar import AnimatedMarker, Bar, BouncingBar, Counter, ETA, \
    AdaptiveETA, FileTransferSpeed, FormatLabel, Percentage, \
    ProgressBar, ReverseBar, RotatingMarker, \
    SimpleProgress, Timer, UnknownLength
```

4.1.2. Lectura de etiquetas y filtrado de datos

El *dataset* incluye la información relevante de las muestras de audio en archivos *.tsv* que describen a cada una de las muestras, indicando su validez (si la grabación corresponde o no con la transcripción) y los datos del usuario que la grabó, incluyendo el género y la edad, que son los componentes relevantes para la clasificación. Hay muestras que no disponen de estos datos, por tanto, no pueden ser utilizadas en el estudio y tienen que ser eliminadas del conjunto antes de comenzar a trabajar con los datos, ya que la conversión a *.wav* de la muestras y la posterior extracción de características son tareas costosas desde el punto de vista computacional.

Como se ha expuesto anteriormente, el conjunto con los metadatos de las muestras de audio está dividido en cuatro subconjuntos creados por *Mozilla* para tareas de *Machine Learning* [Sección 2.3.3.1], pero la división realizada se basa en criterios que no son de utilidad directa para el estudio, por lo que se han combinado todos los conjuntos de metadatos y se han eliminado las muestras inválidas en función del criterio de exclusión expuesto al comienzo de este apartado.

4.1.2.1. Conjunto train

```
train_df = pd.read_csv('en/train.tsv', sep='\t')
train_df_final = train_df[train_df['gender'].notna()]
train_df_final = train_df_final[train_df_final['age'].notna()]
train_df_final.to_csv('train_no_na.tsv', sep='\t', index=False)
```

4.1.2.2. Conjunto dev

```
dev_df = pd.read_csv('en/dev.tsv', sep='\t')
dev_df_final = dev_df[dev_df['gender'].notna()]
dev_df_final = dev_df_final[dev_df_final['age'].notna()]
dev_df_final.to_csv('dev_no_na.tsv', sep='\t', index=False)
```

4.1.2.3. Conjunto test

```
test_df = pd.read_csv('en/test.tsv', sep='\t', dtype={'locale': str, 'segment': str})
test_df_final = test_df[test_df['gender'].notna()]
test_df_final = test_df_final[test_df_final['age'].notna()]
test_df_final.to_csv('test_no_na.tsv', sep='\t', index=False)
```

4.1.2.4. Conjunto valid

```
valid_df = pd.read_csv('en/validated.tsv', sep='\t')
valid_df_final = valid_df[valid_df['gender'].notna()]
valid_df_final = valid_df_final[valid_df_final['age'].notna()]
valid_df_final.to_csv('valid_no_na.tsv', sep='\t', index=False)
```

4.1.2.5. Conjunto invalid

```
invalid_df = pd.read_csv('en/invalidated.tsv', sep='\t', dtype={'locale': str, 'segment': str})
invalid_df_final = invalid_df[invalid_df['gender'].notna()]
invalid_df_final = invalid_df_final[invalid_df_final['age'].notna()]
invalid_df_final.to_csv('invalid_no_na.tsv', sep='\t', index=False)
```

4.1.2.6. Conjunto other

```
other_df = pd.read_csv('en/other.tsv', sep='\t', dtype={'locale': str, 'segment': str})
other_df_final = other_df[other_df['gender'].notna()]
other_df_final = other_df_final[other_df_final['age'].notna()]
other_df_final.to_csv('other_no_na.tsv', sep='\t', index=False)
```

4.1.3. Combinación de conjuntos de datos con información de las muestras

Una vez se han eliminado las muestras no deseadas de cada subconjunto, se combinan para formar un solo *DataFrame* que contenga todos los datos.

```
audio_metadata_df = pd.concat([train_df, dev_df, test_df, valid_df, invalid_df, other_df])
```

A continuación se muestra un ejemplo de los metadatos que se incluyen para una muestra de voz.

```
client_id:  cb314f70a5f16bebd496ab770f759e3f34682708708cfc...
path:       common_voice_en_18603187.mp3
sentence:   Oh, I was terrified!
up_votes:   2
down_votes: 1
age:        fourties
gender:     female
accent:     NaN
locale:     en
segment:    NaN
```

Al haber combinado los subconjuntos, es necesario comprobar que no se ha incluido ninguna muestra duplicada que pudiera estar contenida en varios subconjuntos a la vez.

```
audio_metadata_df.duplicated(keep='first').sum()
```

```
436138
```

Se observa que hay 436138 muestras duplicadas en el conjunto final, por lo que se eliminan.

```
audio_metadata_no_duplicates_df = audio_metadata_df.drop_duplicates(keep='first')
```

El metadato principal para el estudio que se requiere de las muestras es la edad. En el *dataset* se distinguen nueve grupos de edad, tal como se indica en el siguiente fragmento de código.

```
audio_metadata_no_duplicates_df.age.unique()
```

```
array(['fourties', 'twenties', 'thirties', 'fifties', 'seventies', 'teens', 'eighties', 'sixties',
'nineties'], dtype=object)
```

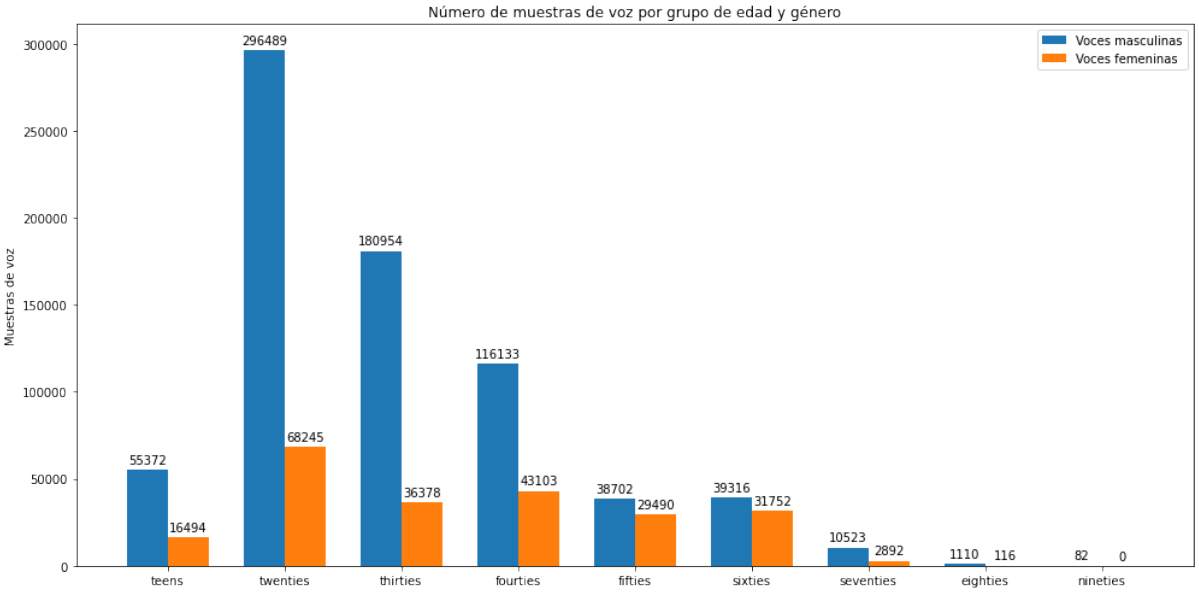
4.1.4. Ajuste de dimensiones del conjunto de datos

Debido a la extensión del conjunto de datos, como se puede comprobar en el fragmento de código siguiente y en la Figura 4.2, es necesario seleccionar una cantidad de muestras proporcional de cada uno de los grupos de edad, con el fin de reducir el tiempo de entrenamiento de los modelos de clasificación.

```
for x in audio_metadata_no_duplicates_df.age.unique():
    print(x)
    print('male' + '\t' +
    str((audio_metadata_no_duplicates_df.loc[(audio_metadata_no_duplicates_df['age'] == x) &
    (audio_metadata_no_duplicates_df['gender'] == 'male')].count()).path))
    print('female' + '\t' +
    str((audio_metadata_no_duplicates_df.loc[(audio_metadata_no_duplicates_df['age'] == x) &
    (audio_metadata_no_duplicates_df['gender'] == 'female')].count()).path))
    print('total' + '\t' +
    str((audio_metadata_no_duplicates_df.loc[(audio_metadata_no_duplicates_df['age'] == x)]
    .count()).path))
    print('-----------------------------------')
```

```
fourties
male    116133
female  43103
total   160233
-----------------------------------
twenties
male    296489
female  68245
total   371815
-----------------------------------
thirties
male    180954
female  36378
total   220928
-----------------------------------
fifties
male    38702
female  29490
total   68259
-----------------------------------
seventies
male    10523
female  2892
total   13415
-----------------------------------
teens
male    55372
female  16494
total   94091
-----------------------------------
eighties
male    1110
female  116
total   1247
-----------------------------------
sixties
male    39316
female  31752
total   71126
-----------------------------------
nineties
male    82
female  0
total   127
-----------------------------------
```

Figura 4.2: Distribución de las muestras del conjunto *Common Voice* por edad y género



Dentro del *dataset* se incluyen 71866 entradas de voz para el grupo de edad `teens` [Figura 4.2]. Para tener grupos de datos proporcionados, se ha utilizado una cantidad similar de datos de los otros grupos de edad. Como se puede observar en la Figura 4.2, los grupos de edad de 70, 80 y 90 años no contienen suficientes datos para poder obtener una muestra representativa. El conjunto de datos que se ha seleccionado del total incluye 5000 hombres y 5000 mujeres para cada grupo de edad, con un total de 60000 muestras.

```
for x in audio_metadata_no_duplicates_df.age.unique():
    if x == 'teens':
        audio_metadata_teens_male = audio_metadata_no_duplicates_df
        .loc[(audio_metadata_no_duplicates_df['age'] == x) &
        (audio_metadata_no_duplicates_df['gender'] == 'male')]
        .sample(n=5000, replace=False, random_state=1)
        audio_metadata_teens_female = audio_metadata_no_duplicates_df
        .loc[(audio_metadata_no_duplicates_df['age'] == x) &
        (audio_metadata_no_duplicates_df['gender'] == 'female')]
        .sample(n=5000, replace=False, random_state=1)

        print(x)
        print('male:' + str(audio_metadata_teens_male.path.count()))
        print('female:' + str(audio_metadata_teens_female.path.count()))
        print('____________________________________')
    elif x == 'twenties':
        audio_metadata_twenties_male = audio_metadata_no_duplicates_df
        .loc[(audio_metadata_no_duplicates_df['age'] == x) &
        (audio_metadata_no_duplicates_df['gender'] == 'male')]
        .sample(n=5000, replace=False, random_state=1)
        audio_metadata_twenties_female = audio_metadata_no_duplicates_df
        .loc[(audio_metadata_no_duplicates_df['age'] == x) &
        (audio_metadata_no_duplicates_df['gender'] == 'female')]
        .sample(n=5000, replace=False, random_state=1)

        print(x)
        print('male:' + str(audio_metadata_twenties_male.path.count()))
        print('female:' + str(audio_metadata_twenties_female.path.count()))
        print('____________________________________')
    elif x == 'thirties':
        audio_metadata_thirties_male = audio_metadata_no_duplicates_df
```

```
        .loc[(audio_metadata_no_duplicates_df['age'] == x) &
        (audio_metadata_no_duplicates_df['gender'] == 'male')]
        .sample(n=5000, replace=False, random_state=1)
        audio_metadata_thirties_female = audio_metadata_no_duplicates_df
        .loc[(audio_metadata_no_duplicates_df['age'] == x) &
        (audio_metadata_no_duplicates_df['gender'] == 'female')]
        .sample(n=5000, replace=False, random_state=1)

        print(x)
        print('male:' + str(audio_metadata_thirties_male.path.count()))
        print('female:' + str(audio_metadata_thirties_female.path.count()))
        print('_____________________________________')
    elif x == 'fourties':
        audio_metadata_fourties_male = audio_metadata_no_duplicates_df
        .loc[(audio_metadata_no_duplicates_df['age'] == x) &
        (audio_metadata_no_duplicates_df['gender'] == 'male')]
        .sample(n=5000, replace=False, random_state=1)
        audio_metadata_fourties_female = audio_metadata_no_duplicates_df
        .loc[(audio_metadata_no_duplicates_df['age'] == x) &
        (audio_metadata_no_duplicates_df['gender'] == 'female')]
        .sample(n=5000, replace=False, random_state=1)

        print(x)
        print('male:' + str(audio_metadata_fourties_male.path.count()))
        print('female:' + str(audio_metadata_fourties_female.path.count()))
        print('_____________________________________')
    elif x == 'fifties':
        audio_metadata_fifties_male = audio_metadata_no_duplicates_df
        .loc[(audio_metadata_no_duplicates_df['age'] == x) &
        (audio_metadata_no_duplicates_df['gender'] == 'male')]
        .sample(n=5000, replace=False, random_state=1)
        audio_metadata_fifties_female = audio_metadata_no_duplicates_df
        .loc[(audio_metadata_no_duplicates_df['age'] == x) &
        (audio_metadata_no_duplicates_df['gender'] == 'female')]
        .sample(n=5000, replace=False, random_state=1)

        print(x)
        print('male:' + str(audio_metadata_fifties_male.path.count()))
        print('female:' + str(audio_metadata_fifties_female.path.count()))
        print('_____________________________________')
    elif x == 'sixties':
        audio_metadata_sixties_male = audio_metadata_no_duplicates_df
        .loc[(audio_metadata_no_duplicates_df['age'] == x) &
        (audio_metadata_no_duplicates_df['gender'] == 'male')]
        .sample(n=5000, replace=False, random_state=1)
        audio_metadata_sixties_female = audio_metadata_no_duplicates_df
        .loc[(audio_metadata_no_duplicates_df['age'] == x) &
        (audio_metadata_no_duplicates_df['gender'] == 'female')]
        .sample(n=5000, replace=False, random_state=1)

        print(x)
        print('male:' + str(audio_metadata_sixties_male.path.count()))
        print('female:' + str(audio_metadata_sixties_female.path.count()))
        print('_____________________________________')
    else:
        pass
```

4.1.5. Selección de archivos de audio

Una vez se tiene el conjunto de muestras final, hay que seleccionar los archivos de audio para convertirlos a *.wav*. Como se ha visto anteriormente, cada una de las muestras con metadatos indican en la variable `path` a que archivo de audio pertenecen, por lo que para seleccionar los archivos de audio que se han de convertir se obtiene una lista con los valores de la variable `path` de todas las muestras, que se comparará con la lista de nombres de archivos disponibles en el *dataset*. Realizando la intersección de ambas listas se obtendrán la lista de nombres de los archivos con las muestras de audio que corresponden a los metadatos seleccionados. Para acelerar el procesamiento en la comparación de los nombres de los archivos, se ordenarán previamente a su comparación y se transformará cada lista en un `Set`.

```
filenames = []
# Listado de nombres de ficheros con muestras de audio
filenames = os.listdir('en/clips/')
filenames.sort()

# Listado de ficheros seleccionados
filenames_selection = audio_metadata_final.path.tolist()
filenames_selection.sort()

# Intersección de ambas listas
filenames_final = list(set(filenames) & set(filenames_selection))
```

4.1.6. Transformación de archivos de audio a formato .wav

Teniendo el conjunto de datos final, se transforman los archivos de audio a formato sin compresión para extraer las características con *openSMILE*.

```
def convert_files(directory, filenames):
    size = len(filenames)

    # Inicialización de la barra de carga para mostrar el estado del proceso
    widgets = [Percentage(),
               ' ', Bar(),
               ' ', ETA()]
    pbar = ProgressBar(widgets=widgets, maxval=size)
    pbar.start()

    i = 0
    for x in filenames:
      try:
        # Lectura del fichero .mp3
        sound = AudioSegment.from_mp3(directory + x)

        # Se separa el nombre y la extensión del fichero
        name = os.path.splitext(x)[0]
        name += '.wav'
        sound.export(name, format="wav")

        # Actualización de la barra de carga
        i += 1
        pbar.update(i)
      except pydub.exceptions.CouldntDecodeError:
        print('Ha ocurrido un error en el archivo ' + x)

    pbar.finish()
```

4.1.7. Extracción de características

Para extraer las características de los archivos de audio, es necesario cargar el conjunto de características que se van a usar. En este caso, se recurre al conjunto *eGEMAPSv02*, disponible de forma nativa en *openSMILE*.

```
smile = opensmile.Smile(
    feature_set=opensmile.FeatureSet.eGeMAPSv02,
    feature_level=opensmile.FeatureLevel.Functionals)
```

Teniendo el conjunto de características configurado, se procede a la extracción de las características de las muestras, leyendo cada uno de los archivos y almacenando el resultado en un *array*.

```
features = []
paths = []
audio_files = []

audio_files = os.listdir('wav/')
size = len(audio_files)

# Inicialización de la barra de carga
widgets = [Percentage(),
               ' ', Bar(),
               ' ', ETA()]
pbar = ProgressBar(widgets=widgets, maxval=size)
pbar.start()

i = 0
for x in audio_files:
    # Lectura del fichero con la muestra de audio
    signal, sampling_rate = audiofile.read('wav/' + x, always_2d=True)

    # Extracción de características con openSMILE
    features.append(smile.process_signal(signal, sampling_rate))

    # Extracción del nombre del fichero de audio
    paths.append((os.path.splitext(x))[0])

    # Actualización de la barra de carga
    i += 1
    pbar.update(i)

pbar.finish()
```

Para poder combinar las características extraídas de las muestras con el resto de metadatos, es necesario convertir el *array* de características en un *DataFrame*. Para ello se recorren todas las posiciones del *arrray* para extraer los valores y guardarlos en un *DataFrame*.

```
size = len(features)
features_df = pd.DataFrame(features[0])

# Inicialización de la barra de carga
widgets = [Percentage(),
               ' ', Bar(),
               ' ', ETA()]
pbar = ProgressBar(widgets=widgets, maxval=size)
pbar.start()

for i in range(1, size):
  features_df = features_df.append(features[i])

  # Actualización de la barra de carga
  pbar.update(i)

pbar.finish()
```

Teniendo la estructura de datos adecuada, se añade una nueva columna al *DataFrame*, que indica el nombre del fichero al que corresponden las características. Los nombres de los ficheros almacenados en la variable `paths` permitirán combinar los metadatos de las muestras con las características extraídas con *openSMILE*.

```
features_df.insert(0, 'path', paths)
```

4.1.8. Combinación de metadatos y características de las muestras

Para unir los *DataFrames* de metadatos y características se ordenarán las filas por nombre, utilizando la columna `path`, común a ambos conjuntos, para posicionar en orden adecuado los datos, haciendo que cada una de las filas de los *DataFrames* haga referencia al mismo archivo de audio.

Antes de ordenar las filas del *DataFrame* de metadatos por nombre de fichero, es necesario aplicar una transformación a la columna `path`, donde se almacenan los nombres de fichero, para separar la extensión del resto del nombre del archivo y eliminarla. Para llevar a cabo esta tarea, se define una función `extract_name(...)`, que aplicará la separación de nombre y extensión a cada valor en la columna. Para aplicar esta función a toda una columna, se utiliza la función `apply(...)` de la librería *pandas*.

```
def extract_name(x):
    # Separa el nombre del archivo de la extensión, eliminando esta última
    return x.split('.')[0]
```

El último paso para obtener la unión de *DataFrames* es, como se ha expuesto anteriormente, ordenar las filas por orden alfabético en función del nombre de archivo indicado en la variable `path` y combinar los *DataFrames* con la función `join(...)` de *pandas*.

```
audio_metadata_final_df['path'] = audio_metadata_final_df['path'].apply(extract_name)
audio_metadata_final_df = audio_metadata_final_df.sort_values('path', ignore_index=True)
```

```
features_path_df = features_path_df.sort_values('path', ignore_index=True)
# Se elimina la columna path, que está incluida en el DataFrame de metadatos
features_path_df = features_path_df.drop(labels='path', axis=1)
```

```
# Combinación de los dataframes de metadatos y características
audio_features_df = audio_metadata_final_df.join(features_path_df)
```

Como los datos están ordenados por nombre de archivo, es necesario generar un orden aleatorio de datos para asegurar que los conjuntos de entrenamiento y validación que se obtengan a partir de ellos incluyan muestras representativas de los datos que permitan resultados coherentes.

```
audio_features_random_df = audio_features_df.sample(frac=1, random_state=1).reset_index(drop=True)
```

4.2. Desarrollo del modelo de clasificación de voz

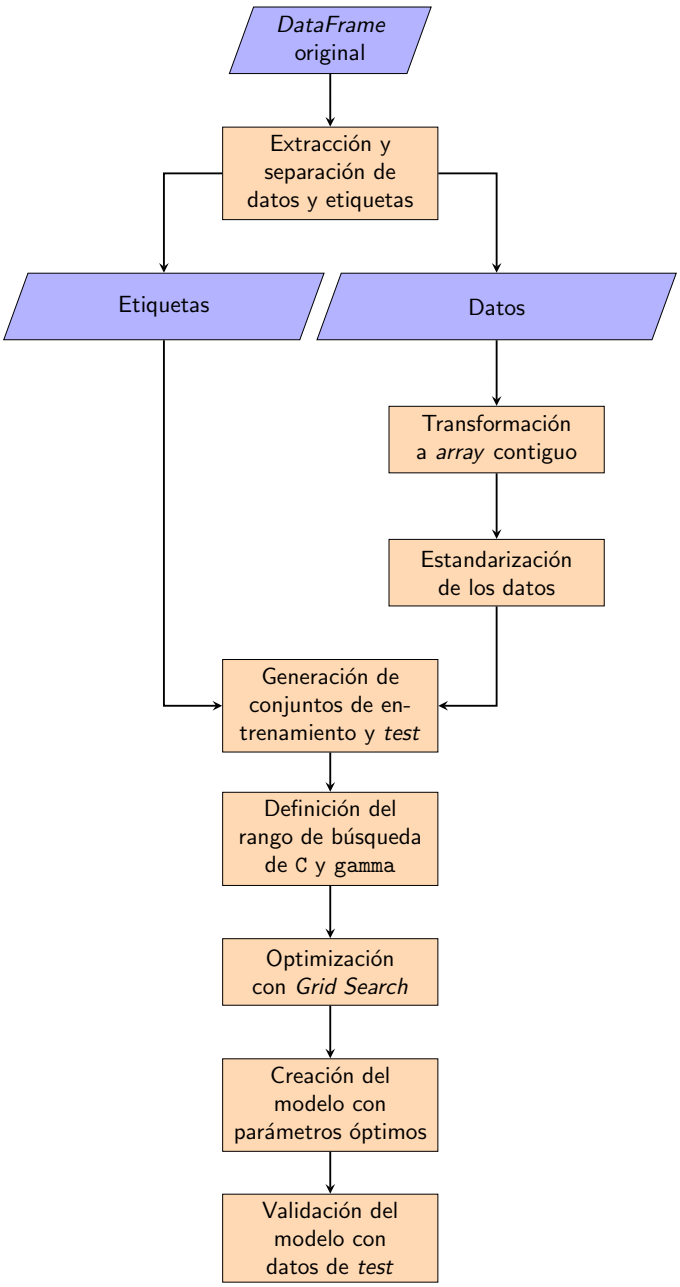
Para el desarrollo del modelo de voz se ha hecho uso de la librería *Scikit Learn* [Sección 2.4.2]. Para comparar el rendimiento de los modelos en idiomas diferentes, se han desarrollado cuatro modelos, dos de ellos para clasificación de voz con el conjunto de datos en inglés y otros dos modelos para clasificación con el conjunto de datos en español. Al tener dimensiones diferentes, los conjuntos de datos también han permitido evaluar como afecta el tamaño del *dataset* al rendimiento de los modelos.

Los métodos escogidos para realizar la clasificación de voz han sido *Support Vector Machines* y *Random Forest*, ambos de la rama de aprendizaje supervisado, al disponer de las etiquetas con la edad en todas las muestras. Adicionalmente, se realizaron pruebas con modelos estadísticos (*Gaussian Mixture Models*) para comprobar el rendimiento de alternativas no supervisadas. Al ser un método de aprendizaje no supervisado, son necesarios una serie de ajustes en el modelo de clasificación para convertir la herramienta de *clustering* en un método que permita operar con datos etiquetados, como se detalla en el apartado correspondiente.

4.2.1. Clasificación con Support Vector Machines

La clasificación de voz con modelos de la familia SVM sigue un esquema clásico, como se muestra en la Figura 4.3. A partir de las características de las muestras de voz extraídas del conjunto *Common Voice*, se separan las etiquetas de los propios datos y se realiza el procesamiento pertinente de los datos antes de pasar al ajuste de los modelos. La optimización se realiza con la funcionalidad *Grid Search* para definir los parámetros óptimos del modelo y validarlo con el conjunto de datos reservados para *test*. Los apartados siguientes tratan en detalle el proceso expuesto en la figura.

Figura 4.3: Diagrama de desarrollo del clasificador con SVM



4.2.1.1. Carga de librerías

Scikit Learn proporciona todas las herramientas necesarias para trabajar con los datos, desde la normalización y creación de conjuntos de entrenamiento y *test* hasta la visualización de resultados para la evaluación de los modelos. Además de las funciones propias de *Scikit Learn*, es necesario cargar librerías para manejar los datos (*Numpy* y *Pandas*) y una librería para crear gráficos en caso de que sean necesarios (*Matplotlib*).

```
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

import seaborn as sns; sns.set()

from sklearn import preprocessing
from sklearn.svm import SVC
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.pipeline import make_pipeline
from sklearn.metrics import classification_report, confusion_matrix
```

4.2.1.2. Preparación del conjunto de datos

Partiendo del *DataFrame* creado en la Sección 4.1, que contiene las características de las muestras de voz junto a los metadatos asociados, es necesario generar dos conjuntos de entrada y de salida, que se han denominado `X` e `Y`. `X` representa el conjunto de las características de las muestras de voz sin incluir los metadatos ni las etiquetas, e `Y` representa el conjunto de las etiquetas de edad de cada una de las muestras.

Para obtener los conjuntos descritos anteriormente de los datos originales se utilizan las funciones de manipulación de *DataFrames* que proporciona *pandas*.

```
audio_features = pd.read_csv('audio_features_random.csv')
audio_features
```

El conjunto `X` se extrae del *DataFrame* `audio_features` eliminando todas las columnas de metadatos, para dejar únicamente aquellas correspondientes a las características del audio que se extrajeron con *openSMILE*.

```
X = audio_features.drop(['client_id', 'path', 'sentence', 'up_votes', 'down_votes',
'age', 'gender','accent'], axis=1)
```

La implementación de SVM en *Scikit Learn* requiere que los datos de entrada a ciertos métodos estén almacenados en *arrays* contiguos en memoria ordenados por columnas y contengan datos de tipo *float* de 32 *bits* [34]. Si no siguen estas condiciones, los datos serán copiados antes de cada operación, lo que ralentiza el funcionamiento del algoritmo. La librería *numpy* contiene una función para transformar *arrays* no contiguos en contiguos, por lo que antes de continuar con el tratamiento de los datos se realiza la transformación.

```
X = np.ascontiguousarray(X, dtype=np.float32)
```

La extracción de las etiquetas del *DataFrame* original sigue un proceso similar. La columna de datos age del *DataFrame*, que contiene todos las etiquetas de edad, se separa del resto de datos y se almacena en un *array*.

```
Y = audio_features['age'].to_numpy()
```

Una vez se tienen los conjuntos de datos definidos, es necesario realizar una división en el *dataset* para obtener los conjuntos de entrenamiento y *test*. Para ello, se ha optado por utilizar un 80 % de los datos para el entrenamiento del modelo y reservar un 20 % para la validación.

```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.20, random_state=2)
```

4.2.1.3. Entrenamiento y optimización del modelo

Para realizar la búsqueda de parámetros óptimos se ha utilizado la funcionalidad *Grid Search* incluida en la librería *Scikit Learn*. Está función permite evaluar el rendimiento del modelo aplicando combinaciones de los diferentes parámetros que ajustan el modelo a través de un proceso de validación cruzada.

Al no tener información a priori sobre la distribución de los datos, se ha escogido una función de *kernel* radial o *Radial Basis Function* (RBF) para el modelo. Para este tipo de función de *kernel* es necesario ajustar dos parámetros, `C` y `gamma`.

El parámetro `C` ajusta el balance entre la clasificación errónea de los datos de entrenamiento y la complejidad de la función de decisión a través del margen. Conforme se aumente el valor de `C`, el modelo intentará clasificar correctamente el mayor número de ejemplos de entrenamiento posible reduciendo el margen, pero aumentará la complejidad de la función de decisión, creciendo también el tiempo de entrenamiento. El valor adecuado de `C` dependerá del conjunto de datos, y será aquel que permita tener el mayor margen (menor complejidad de la función de decisión) y la mayor tasa de clasificación correcta de ejemplos.

Con el parámetro `gamma` se define la influencia de cada ejemplo en la función de decisión. Cuanto más alto sea el valor de `gamma`, menor será la zona de influencia de los vectores de soporte, y cuanto menor sea `gamma`, menor será el ajuste de la función de decisión a la distribución de los datos.

Al ser parámetros críticos para asegurar el buen funcionamiento del modelo, tal como se indica en la documentación de la librería, para realizar una búsqueda óptima de parámetros se ha utilizado saltos logarítmicos en los valores de `C` y `gamma`. Para `C` se ha escogido un rango entre 0.1 y 1000, mientras que para `gamma` se ha seleccionado un rango entre 0.001 y 0.1.

```
param_grid = {'svc__C': [0.1,1,10,100,1000], 'svc__kernel':['rbf'], 'svc__gamma':[0.1,0.01,0.001]}
```

Los algoritmos de SVM son sensibles a la escala de los datos, por lo que antes de comenzar el entrenamiento de los modelos, se realiza una estandarización independiente de cada una de las características de los datos siguiendo la expresión 4.1 [50]:

$$z = \frac{x - u}{s} \tag{4.1}$$

donde:

x es el dato de entrada

u es la media de los datos de entrada

s es la desviación estándar de los datos de entrada

Para agrupar las operaciones de estandarización de los datos y el entrenamiento y validación de los modelos SVM, se crea un *pipeline* con las funciones de estandarización y el proceso de *Grid Search*.

```
clf = make_pipeline(StandardScaler(), SVC(random_state=0))
svm_grid_search = GridSearchCV(estimator=clf, param_grid=param_grid, cv=3, verbose=2)
svm_grid_search.fit(X_train, y_train)
```

```
Fitting 3 folds for each of 15 candidates, totalling 45 fits
[...]
GridSearchCV(cv=3,
             estimator=Pipeline(steps=[('standardscaler', StandardScaler()),
                                       ('svc', SVC(random_state=0))]),
             param_grid={'svc__C': [0.1, 1, 10, 100, 1000],
                         'svc__gamma': [0.1, 0.01, 0.001],
                         'svc__kernel': ['rbf']},
             verbose=2)
```

4.2.1.4. Resultados en el conjunto Common Voice en inglés

Como se mostró en la Sección 2.3.3.1, el conjunto de datos en inglés es el más amplio de los dos, lo que ha permitido realizar pruebas con dos tamaños del conjunto de datos diferentes. En la primera prueba se ha trabajado con una versión del conjunto con 18000 muestras, de tamaño similar al *dataset* en español y en la segunda se ha optado por una muestra de tamaño extendido, que contiene 60000 muestras.

Resultados con el conjunto de 18000 muestras Siguiendo con el fragmento de código anterior, se muestran los parámetros óptimos para el modelo SVM de clasificación con 18000 muestras.

```
svm_grid_search.best_params_
```

```
{'svc__C': 10, 'svc__gamma': 0.1, 'svc__kernel': 'rbf'}
```

Como se puede observar en el extracto superior, los mejores parámetros obtenidos para este conjunto de datos son:

- $C = 10$
- gamma $= 0.1$

Ambos valores se encuentra dentro de los límites del rango de búsqueda seleccionado, lo que indica que son valores óptimos para el modelo.

Para evaluar el rendimiento del modelo óptimo obtenido con *Grid Search* se ha hecho uso de la utilidad *Classification Report* [51], que aporta métricas acerca del funcionamiento del modelo para cada una de las clases y a nivel global, además de la representación de la matriz de confusión.

```
clf = svm_grid_search.best_estimator_
clf.fit(X_train, y_train)
y_pred_test = clf.predict(X_test)
print(classification_report(y_test, y_pred_test))
```

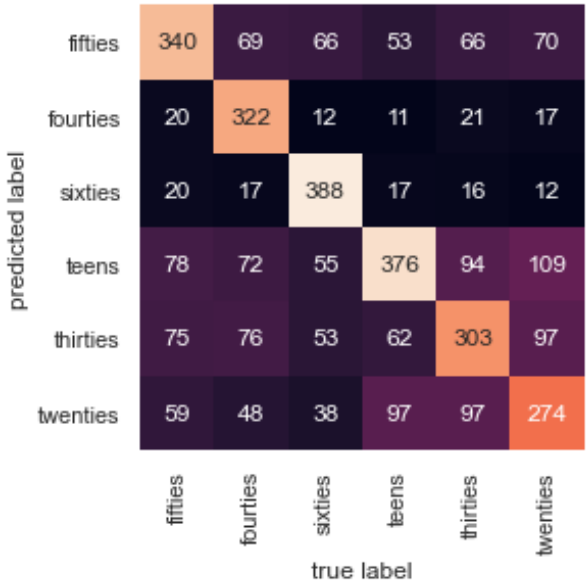
```
              precision    recall  f1-score   support

     fifties       0.51      0.57      0.54       592
    fourties       0.80      0.53      0.64       604
     sixties       0.83      0.63      0.72       612
       teens       0.48      0.61      0.54       616
    thirties       0.45      0.51      0.48       597
    twenties       0.45      0.47      0.46       579

    accuracy                           0.56      3600
   macro avg       0.59      0.56      0.56      3600
weighted avg       0.59      0.56      0.56      3600
```

```
mat = confusion_matrix(y_test, y_pred_test)
sns.heatmap(mat.T, square=True, annot=True, fmt='d', cbar=False,
xticklabels=['fifties', 'fourties', 'sixties', 'teens', 'thirties', 'twenties'],
yticklabels=['fifties', 'fourties', 'sixties', 'teens', 'thirties', 'twenties'])
plt.xlabel('true label')
plt.ylabel('predicted label');
```

Figura 4.4: Matriz de confusión del clasificador SVM con 18000 muestras. *Common Voice* en inglés



Resultados con el conjunto extendido de 60000 muestras Para el conjunto extendido del *dataset* en inglés se han obtenido los siguientes parámetros.

```
svm_grid_search.best_params_
```

```
{'svc__C': 1000, 'svc__gamma': 0.1, 'svc__kernel': 'rbf'}
```

En el fragmento superior se muestran los parámetros óptimos calculados para el modelo de clasificación:

- $C = 1000$
- gamma $= 0.1$

Como en el caso anterior, se utiliza la funcionalidad *Classification Report* y la matriz de confusión para evaluar el rendimiento del modelo.

```
clf = svm_grid_search.best_estimator_
clf.fit(X_train, y_train)
y_pred_test = clf.predict(X_test)
print(classification_report(y_test, y_pred_test))
```

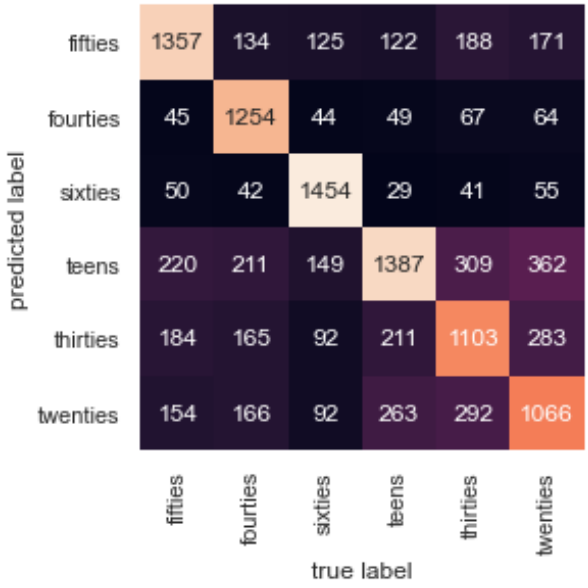
```
              precision    recall  f1-score   support

     fifties       0.65      0.68      0.66      2010
    fourties       0.82      0.64      0.72      1972
     sixties       0.87      0.74      0.80      1956
       teens       0.53      0.67      0.59      2061
    thirties       0.54      0.55      0.55      2000
    twenties       0.52      0.53      0.53      2001

    accuracy                           0.64     12000
   macro avg       0.66      0.64      0.64     12000
weighted avg       0.65      0.64      0.64     12000
```

```
mat = confusion_matrix(y_test, y_pred_test)
sns.heatmap(mat.T, square=True, annot=True, fmt='d', cbar=False,
xticklabels=['fifties', 'fourties', 'sixties', 'teens', 'thirties', 'twenties'],
yticklabels=['fifties', 'fourties', 'sixties', 'teens', 'thirties', 'twenties'])
plt.xlabel('true label')
plt.ylabel('predicted label');
```

Figura 4.5: Matriz de confusión del clasificador SVM con 60000 muestras. *Common Voice* en inglés



4.2.1.5. Resultados en el conjunto Common Voice en español

El conjunto de datos en español cuenta con aproximadamente 18000 muestras. Después de realizar la búsqueda de parámetros óptimos con *Grid Search* para este conjunto de datos, los parámetros óptimos del modelo son:

```
svm_grid_search.best_params_
```

```
{'svc__C': 10, 'svc__gamma': 0.01, 'svc__kernel': 'rbf'}
```

Para el *dataset* en español, los mejores parámetros obtenidos son:

- $C = 10$
- gamma $= 0.01$

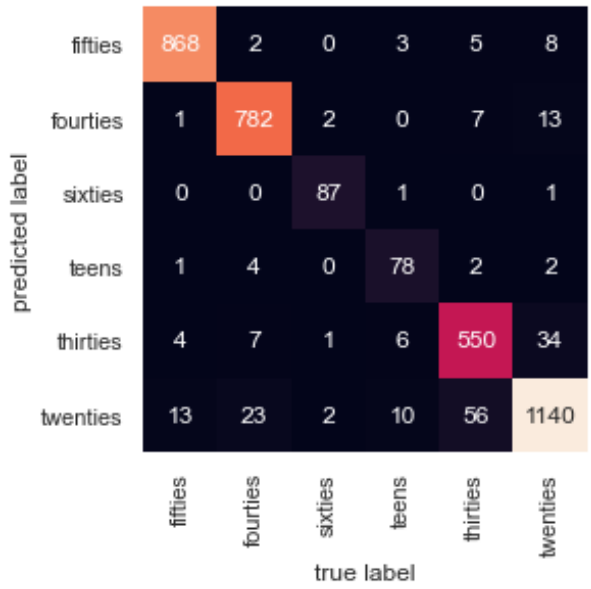
```
clf = svm_grid_search.best_estimator_
clf.fit(X_train, y_train)
y_pred_test = clf.predict(X_test)
print(classification_report(y_test, y_pred_test))
```

Para presentar los datos acerca del rendimiento del modelo se utiliza de nuevo la función *Classification Report* y la matriz de confusión.

	precision	recall	f1-score	support
fifties	0.98	0.98	0.98	887
fourties	0.97	0.96	0.96	818
sixties	0.98	0.95	0.96	92
teens	0.90	0.80	0.84	98
thirties	0.91	0.89	0.90	620
twenties	0.92	0.95	0.93	1198
accuracy			0.94	3713
macro avg	0.94	0.92	0.93	3713
weighted avg	0.94	0.94	0.94	3713

```
mat = confusion_matrix(y_test, y_pred_test)
sns.heatmap(mat.T, square=True, annot=True, fmt='d', cbar=False,
xticklabels=['fifties', 'fourties', 'sixties', 'teens', 'thirties', 'twenties'],
yticklabels=['fifties', 'fourties', 'sixties', 'teens', 'thirties', 'twenties'])
plt.xlabel('true label')
plt.ylabel('predicted label');
```

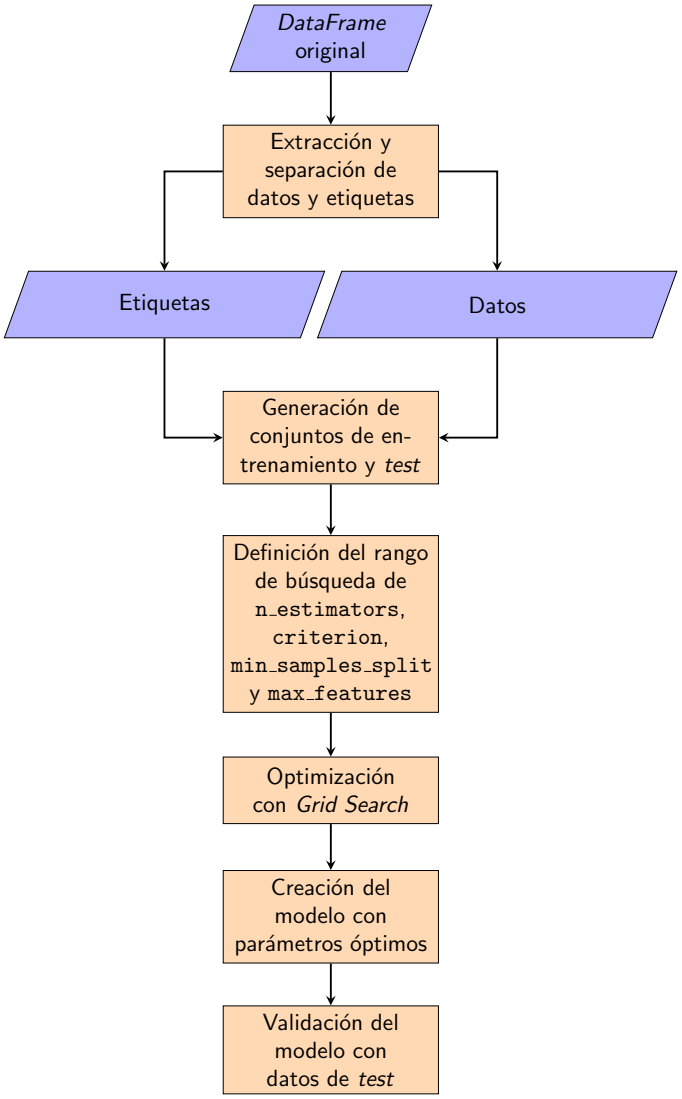
Figura 4.6: Matriz de confusión del clasificador SVM. Common Voice en español



4.2.2. Clasificación con Random Forest

Como se puede observar en la Figura 4.7, el proceso de creación de los modelos de clasificación con *Random Forest* es similar al seguido para los modelos basados en SVM. A partir de los datos extraídos en la fase previa, se generan los conjuntos de entrenamiento y validación y se optimizan los parámetros del modelo con la función *Grid Search*.

Figura 4.7: Diagrama de desarrollo del clasificador con *Random Forest*



4.2.2.1. Carga de librerías

Las herramientas necesarias para trabajar con *Random Forest* en *Scikit Learn* son similares a las expuestas en la Sección 4.2.1. Como en el caso anterior, son necesarias librerías para el manejo de los datos (*Numpy* y *Pandas*) y gráficos (*Matplotlib*), además de la propia *Scikit Learn*.

```
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

import seaborn as sns; sns.set()

from sklearn import preprocessing
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.pipeline import make_pipeline
from sklearn.metrics import classification_report, confusion_matrix
```

4.2.2.2. Preparación del conjunto de datos

Se parte de nuevo del *DataFrame* desarrollado en la sección de extracción de características de las muestras de audio. De igual forma que en la sección anterior, se obtienen los conjuntos `X` e `Y` con los datos de entrada y las etiquetas de salida, respectivamente.

En esta ocasión no es necesario almacenar el *array* con los datos de entrada en una estructura contigua en memoria, ni es necesario estandarizar los datos, ya que los árboles de decisión no son sensibles a la escala de los datos.

```
audio_features = pd.read_csv('audio_features_random.csv')

X = audio_features.drop(['client_id', 'path', 'sentence', 'up_votes', 'down_votes',
'age', 'gender','accent', 'locale', 'segment'], axis=1)

Y = audio_features['age'].to_numpy()
```

Teniendo los conjuntos `X` e `Y` definidos se crean las divisiones correspondientes al entrenamiento y *test*. De nuevo se ha optado por utilizar un 80 % de los datos para entrenamiento y un 20 % de los datos para *test*.

```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.20, random_state=2)
```

4.2.2.3. Entrenamiento y optimización del modelo

Para encontrar los parámetros óptimos del modelo de clasificación con *Random Forest* se hace uso de nuevo de la función *Grid Search*, que permitirá evaluar combinaciones de los diferentes parámetros para encontrar la que ofrece el mejor resultado.

Las dos características principales que es necesario ajustar son el número de árboles o estimadores (`n_estimators`) y el número de características con las que se entrena cada árbol (`max_features`). Cuantos más árboles se utilicen, mejor es el resultado, a costa de aumentar el tiempo de entrenamiento. El número de características que se seleccionen afectará al compromiso entre varianza y sesgo. Si se seleccionan pocas características, la varianza se reduce, pero aumenta el sesgo. Como se sugiere en la documentación de la herramienta, una buena aproximación para problemas de clasificación es utilizar un número de características por árbol igual a la raíz cuadrada del número total de características ($\texttt{max\_features} = \sqrt{\texttt{n\_features}}$) [52].

```
param_grid = {'n_estimators': [300,400,500,600], 'criterion': ['gini', 'entropy'],
'min_samples_split': [2,5,10], 'max_features': ['sqrt', None]}
```

Todos los parámetros serán validados con *Grid Search*.

```
# Inicialización de Random Forest
rf = RandomForestClassifier(n_estimators=500, criterion='entropy', max_depth=None, min_samples_split=2,
max_features='sqrt', random_state=0, n_jobs=-1)

rf_grid_search = GridSearchCV(estimator=rf, param_grid=param_grid, n_jobs=-1, cv=3, verbose=2)

rf_grid_search.fit(X_train, y_train)
```

4.2.2.4. Resultados en el conjunto Common Voice en inglés

A continuación se muestran los parámetros óptimos que se han encontrado para el conjunto de datos en inglés, tanto para el conjunto de tamaño igual al de español como la versión extendida de 60000 muestras.

Resultados con el conjunto de 18000 muestras A partir de la búsqueda de parámetros con *Grid Search* se obtienen los siguientes valores óptimos.

```
rf_grid_search.best_params_
```

```
{'criterion': 'entropy',
 'max_features': 'sqrt',
 'min_samples_split': 2,
 'n_estimators': 600}
```

Los parámetros óptimos para el modelo son:

- Criterio de poda = Entropía
- Número de características por árbol = Raíz cuadrada del total
- Número de muestras necesarias para separar un nodo = 2
- Número de árboles = 600

Como en la evaluación de SVM, se utilizará la herramienta *Classification Report* y la matriz de confusión para evaluar el modelo.

```
rf = rf_grid_search.best_estimator_
rf.fit(X_train, y_train)

y_pred_test = rf.predict(X_test)

print(classification_report(y_test, y_pred_test))
```

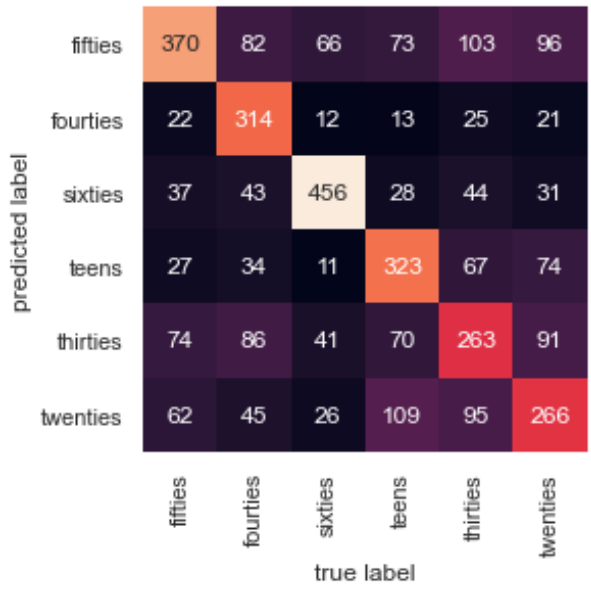
```
              precision    recall  f1-score   support

     fifties       0.47      0.62      0.54       592
    fourties       0.77      0.52      0.62       604
     sixties       0.71      0.75      0.73       612
       teens       0.60      0.52      0.56       616
    thirties       0.42      0.44      0.43       597
    twenties       0.44      0.46      0.45       579

    accuracy                           0.55      3600
   macro avg       0.57      0.55      0.55      3600
weighted avg       0.57      0.55      0.56      3600
```

```
mat = confusion_matrix(y_test, y_pred_test)
sns.heatmap(mat.T, square=True, annot=True, fmt='d', cbar=False, xticklabels=['fifties', 'fourties',
'sixties', 'teens', 'thirties', 'twenties'],
yticklabels=['fifties', 'fourties', 'sixties', 'teens', 'thirties', 'twenties'])
plt.xlabel('true label')
plt.ylabel('predicted label');
```

Figura 4.8: Matriz de confusión del clasificador *Random Forest* con 18000 muestras. Common Voice en inglés



Resultados con el conjunto de 60000 muestras Para el conjunto extendido, los parámetros óptimos que se han obtenido son los siguientes.

```
rf_grid_search.best_params_
```

```
{'criterion': 'entropy',
 'max_features': 'sqrt',
 'min_samples_split': 2,
 'n_estimators': 600}
```

Como se ve en el extracto superior, los parámetros óptimos son:

- Criterio de poda = Entropía
- Número de características por árbol = Raíz cuadrada del total
- Número de muestras necesarias para separar un nodo = 2
- Número de árboles = 600

Se muestra el resultado del *Classification Report* y la matriz de confusión.

```
rf = rf_grid_search.best_estimator_
rf.fit(X_train, y_train)

y_pred_test = rf.predict(X_test)

print(classification_report(y_test, y_pred_test))
```

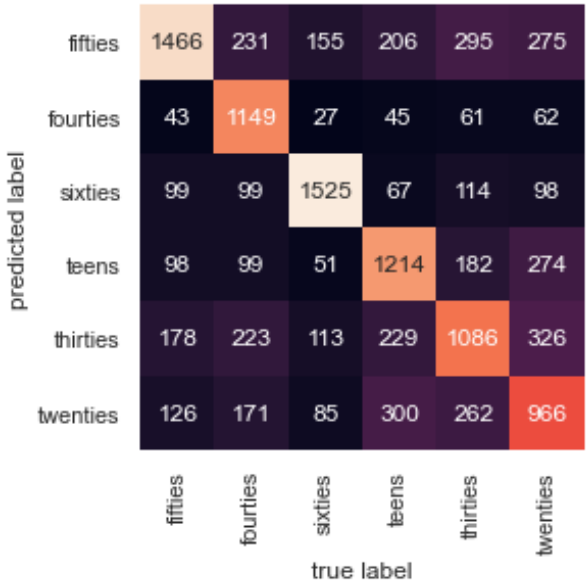
```
              precision    recall  f1-score   support

     fifties       0.56      0.73      0.63      2010
    fourties       0.83      0.58      0.68      1972
     sixties       0.76      0.78      0.77      1956
       teens       0.63      0.59      0.61      2061
    thirties       0.50      0.54      0.52      2000
    twenties       0.51      0.48      0.49      2001

    accuracy                           0.62     12000
   macro avg       0.63      0.62      0.62     12000
weighted avg       0.63      0.62      0.62     12000
```

```
mat = confusion_matrix(y_test, y_pred_test)
sns.heatmap(mat.T, square=True, annot=True, fmt='d', cbar=False, xticklabels=['fifties', 'fourties',
'sixties', 'teens', 'thirties', 'twenties'],
yticklabels=['fifties', 'fourties', 'sixties', 'teens', 'thirties', 'twenties'])
plt.xlabel('true label')
plt.ylabel('predicted label');
```

Figura 4.9: Matriz de confusión del clasificador *Random Forest* con 60000 muestras. Common Voice en inglés



4.2.2.5. Resultados en el conjunto Common Voice en español

El entrenamiento del modelo en español sigue una estructura similar a la descrita en los apartados anteriores. Después de realizar la búsqueda de parámetros óptimos con *Grid Search*, se obtienen los siguientes valores.

```
rf_grid_search.best_params_
```

```
{'criterion': 'entropy',
 'max_features': 'sqrt',
 'min_samples_split': 2,
 'n_estimators': 400}
```

- Criterio de poda = Entropía
- Número de características por árbol = Raíz cuadrada del total
- Número de muestras necesarias para separar un nodo = 2
- Número de árboles = 400

Se calculan las métricas con el rendimiento del modelo y la matriz de confusión.

```
rf = rf_grid_search.best_estimator_
rf.fit(X_train, y_train)

y_pred_test = rf.predict(X_test)

print(classification_report(y_test, y_pred_test))
```

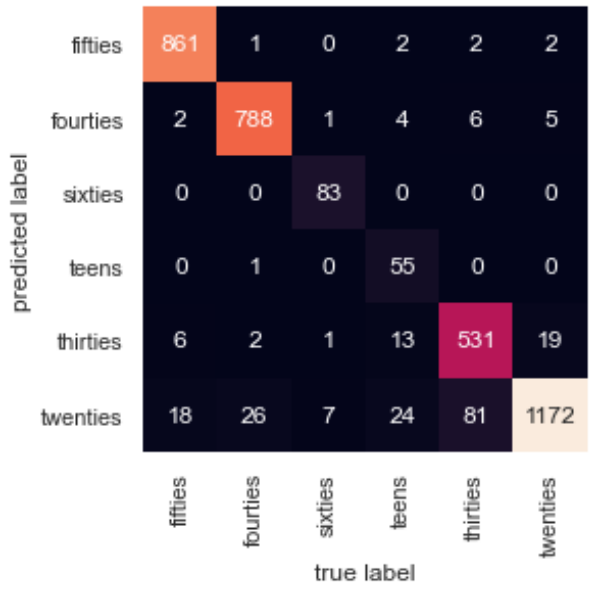
```
              precision    recall  f1-score   support

     fifties       0.99      0.97      0.98       887
    fourties       0.98      0.96      0.97       818
     sixties       1.00      0.90      0.95        92
       teens       0.98      0.56      0.71        98
    thirties       0.93      0.86      0.89       620
    twenties       0.88      0.98      0.93      1198

    accuracy                           0.94      3713
   macro avg       0.96      0.87      0.91      3713
weighted avg       0.94      0.94      0.94      3713
```

```
mat = confusion_matrix(y_test, y_pred_test)
sns.heatmap(mat.T, square=True, annot=True, fmt='d', cbar=False, xticklabels=['fifties', 'fourties',
'sixties', 'teens', 'thirties', 'twenties'],
yticklabels=['fifties', 'fourties', 'sixties', 'teens', 'thirties', 'twenties'])
plt.xlabel('true label')
plt.ylabel('predicted label');
```

Figura 4.10: Matriz de confusión del clasificador *Random Forest*. Common Voice en español

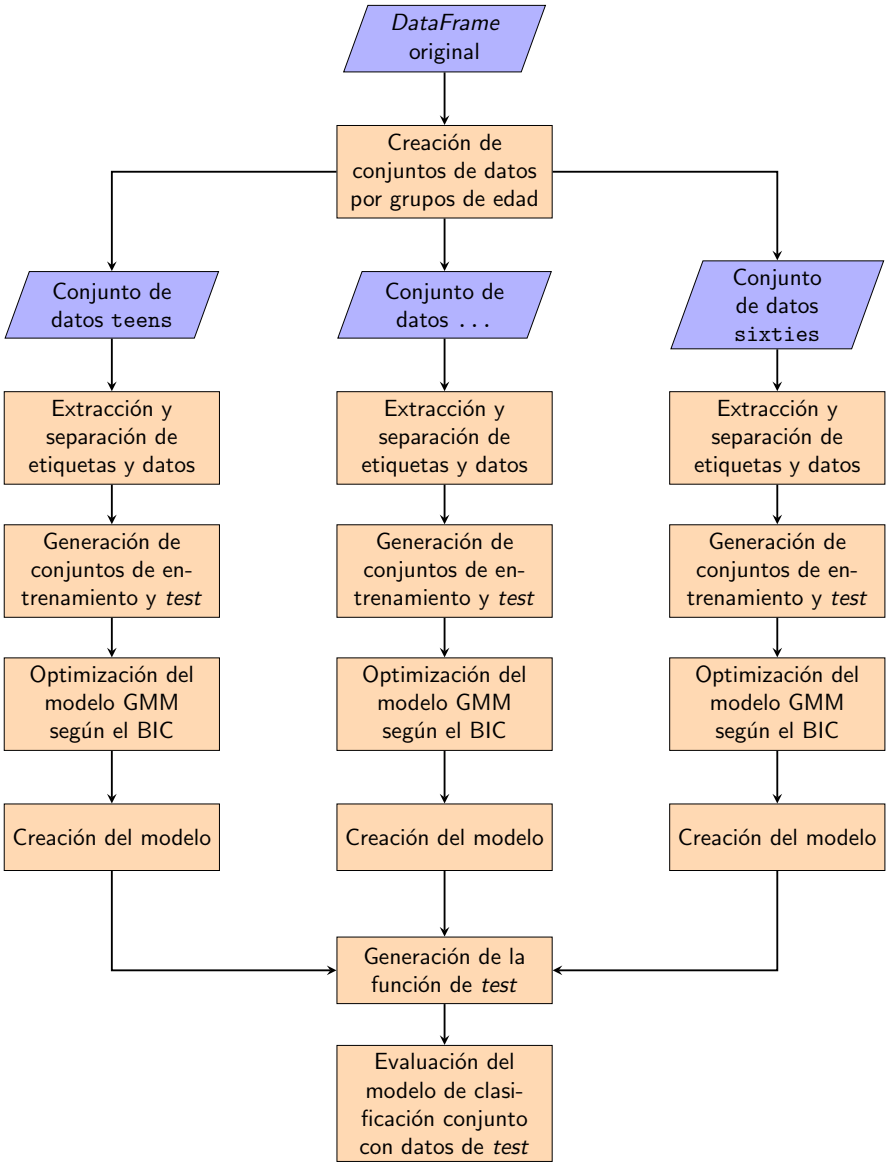


4.2.3. Clasificación con Gaussian Mixture Models

Como se ha explicado al comienzo de la Sección principal 4.2, los modelos estadísticos pertenecen a la familia del aprendizaje no supervisado. Se comportan de un modo similar a los algoritmos de *clustering k-means* pero, mientras que en estos últimos la forma de los *clusters* está predefinida, con los algoritmos *Gaussian Mixture Models* (GMM) se puede trabajar con diferentes funciones que permiten ajustar mejor las agrupaciones a los datos.

El procedimiento para utilizar los modelos GMM en clasificación de voz requiere de varios pasos adicionales con respecto a las anteriores aproximaciones [Figura 4.11]. Como el algoritmo trabaja con datos no etiquetados, el proceso clásico de entrenamiento de un modelo para realizar inferencias aplicado anteriormente no es válido. Es necesario construir un modelo para cada una de las clases, entrenándolos con los datos específicos de cada una de ellas. Las predicciones se realizarán calculando la probabilidad de pertenencia de la muestra a todos los modelos GMM. La etiqueta de salida que se asignará a la muestra será aquella que tengan las muestras con las que se compuso el modelo GMM que arroje la mayor probabilidad de pertenencia.

Figura 4.11: Diagrama de desarrollo del clasificador con GMM



4.2.3.1. Carga de librerías

Se ha hecho uso de las librerías habituales de manejo de datos y gráficos ya descritas, así como de la librería *Scikit Learn* para el desarrollo de los modelos de Inteligencia Artificial.

```
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
import matplotlib as mpl
import pandas as pd
import itertools
from scipy import linalg

import seaborn as sns; sns.set()

from sklearn import preprocessing
from sklearn import mixture
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.pipeline import make_pipeline
from sklearn.metrics import classification_report, confusion_matrix
```

4.2.3.2. Preparación del conjunto de datos

El punto de partida es similar al de los casos estudiados anteriormente. Desde el *DataFrame* original se extraen los conjuntos X e Y con las características de las muestras de audio y las etiquetas, respectivamente.

```
audio_features = pd.read_csv('audio_features_random.csv')

audio_features_data = audio_features.drop(['client_id', 'path', 'sentence',
'up_votes', 'down_votes' , 'age', 'gender','accent', 'locale', 'segment'], axis=1)

X = audio_features_data.to_numpy()
Y = audio_features['age'].to_numpy()
```

4.2.3.3. Entrenamiento y optimización del modelo

Para optimizar de forma general un modelo GMM es necesario ajustar dos parámetros. El primero de ellos es el número de distribuciones que formarán el *mixture* y el segundo es el tipo de covarianza de las clases estimadas, que podrá ser *full*, *spherical*, *tied* o *diagonal* [53]:

- *Full*. Cada componente tiene su propia matriz de covarianza, es decir, cada clúster tendrá su propia forma y posición.
- *Tied*. Todos los componentes comparten la misma matriz de covarianza. Los componentes tendrán la misma forma, que podrá ser de cualquier tipo.
- *Diagonal*. Cada componente tiene su propia matriz diagonal de covarianza. Los ejes que definen a los componentes son paralelos a los ejes de coordenadas, pero cada componente será diferente.
- *Spherical*. Cada componente tiene su propia varianza. Este caso es igual al *diagonal*, pero las matrices de covarianzas serán esféricas.

Con el fin de calcular los parámetros óptimos, se ha utilizado el *Bayesian Information Criterion* (BIC), que ayudará a definir el número de *clusters* óptimos y su covarianza con la función definida en el fragmento de código siguiente (Código disponible en [54]).

```
def evaluate_gmm_model_bic(X, number_components):
    lowest_bic = np.infty
    bic = []
    n_components_range = number_components
    cv_types = ['spherical', 'tied', 'diag', 'full']
    for cv_type in cv_types:
        for n_components in n_components_range:
            # Fit a Gaussian mixture with EM
            gmm = mixture.GaussianMixture(n_components=n_components,
                                          covariance_type=cv_type)
            gmm.fit(X)
            bic.append(gmm.bic(X))
            if bic[-1] < lowest_bic:
                lowest_bic = bic[-1]
                best_gmm = gmm
    bic = np.array(bic)
    color_iter = itertools.cycle(['navy', 'turquoise', 'cornflowerblue', 'darkorange'])
    clf = best_gmm
    bars = []

    plt.figure(figsize=(12, 9))
    spl = plt.subplot(2, 1, 1)
    for i, (cv_type, color) in enumerate(zip(cv_types, color_iter)):
        xpos = np.array(n_components_range) + .2 * (i - 2)
        bars.append(plt.bar(xpos, bic[i * len(n_components_range):
                                  (i + 1) * len(n_components_range)],
                            width=.2, color=color))
    plt.xticks(n_components_range)
    plt.ylim([bic.min() * 1.01 - .01 * bic.max(), bic.max()])
    plt.title('BIC score per model')
    xpos = np.mod(bic.argmin(), len(n_components_range)) + .65 +\
        .2 * np.floor(bic.argmin() / len(n_components_range))
    plt.text(xpos, bic.min() * 0.97 + .03 * bic.max(), '*', fontsize=14)
    spl.set_xlabel('Number of components')
    spl.legend([b[0] for b in bars], cv_types)

    splot = plt.subplot(2, 1, 2)
    Y_ = clf.predict(X)
    for i, (mean, cov, color) in enumerate(zip(clf.means_, clf.covariances_,
                                               color_iter)):
        v, w = linalg.eigh(cov)
        if not np.any(Y_ == i):
            continue
        plt.scatter(X[Y_ == i, 0], X[Y_ == i, 1], .8, color=color)

        angle = np.arctan2(w[0][1], w[0][0])
        angle = 180. * angle / np.pi  # convert to degrees
        v = 2. * np.sqrt(2.) * np.sqrt(v)
        ell = mpl.patches.Ellipse(mean, v[0], v[1], 180. + angle, color=color)
        ell.set_clip_box(splot.bbox)
        ell.set_alpha(.5)
        splot.add_artist(ell)

    plt.xticks(())
    plt.yticks(())
    plt.title(f'Selected GMM: {best_gmm.covariance_type} model, 'f'{best_gmm.n_components} components')
    plt.subplots_adjust(hspace=.35, bottom=.02)
    plt.show()
```

A continuación se divide el proceso para cada uno de los conjuntos de datos, mostrando el ajuste de parámetros para los conjuntos de datos en ingés y español.

4.2.3.4. Resultados en el conjunto Common Voice en inglés

Para generar los conjuntos de entrenamiento y *test* para cada uno de los grupos de edad se opera de forma similar a los casos anteriores, incluyendo un paso de filtrado adicional para escoger sólo aquellos datos que pertenezcan a la clase correspondiente con la que se vaya a entrenar el modelo GMM.

Una vez entrenados todos los modelos GMM correspondientes a los grupos, para evaluar su funcionamiento se utilizan los datos de *test*, que se introducen en cada modelo, para calcular la probabilidad de pertenencia de las muestras a los modelos. Aquel que arroje una mayor probabilidad de pertenencia se selecciona como la clase de la muestra de entrada.

Se muestra un ejemplo para el grupo de edad `teens`, siendo el resto similares, con la única adaptación necesaria a la hora de seleccionar el conjunto, teniendo que cambiar la etiqueta de edad por la del grupo correspondiente.

```
audio_features_teens = audio_features[audio_features['age'] == 'teens']
X_teens = audio_features_teens.drop(['Unnamed: 0', 'client_id', 'path',
'sentence', 'up_votes', 'down_votes' , 'age', 'gender','accent'], axis=1).to_numpy()
Y_teens = audio_features_teens['age'].to_numpy()
X_teens_train, X_teens_test, y_teens_train, y_teens_test = train_test_split(X_teens, Y_teens,
test_size=0.20, random_state=22)
```

Teniendo los conjuntos de entrenamiento y validación preparados para cada grupo de edad, el siguiente paso consiste en entrenar los modelos GMM, uno para grupo edad.

```
evaluate_gmm_model_bic(X_teens_train, range(1,11))

gmm_teens = mixture.GaussianMixture(n_components=9, covariance_type='full', random_state=0)
```

El procedimiento para el resto de grupos de edad es idéntico al mostrado en el fragmento superior.

Para validar los modelos es necesario construir una función de *test* específica, ya que no se está haciendo un uso para *clustering* de los modelos GMM. Como se ha descrito anteriormente, la validación del modelo consiste en evaluar cada muestra de *test* con todos los modelos de los grupos de edad, para obtener la probabilidad de pertenencia de la muestra a cada uno de los GMM. El modelo del grupo de edad que proporcione la probabilidad más alta será considerado como el grupo de edad de la muestra de *test* que se está evaluando.

```
y_age_pred = []
y_prob_list = [0,0,0,0,0,0]

for x in range(0,len(X_age_test)):
    # Calculo de la probabilidad de pertenencia a cada uno de los modelos de los grupos de edad
    y_score_teens = gmm_teens.score(X_age_test[int(x):])
    y_prob_list[0] = y_score_teens

    y_score_twenties = gmm_twenties.score(X_age_test[int(x):])
    y_prob_list[1] = y_score_twenties

    y_score_thirties = gmm_thirties.score(X_age_test[int(x):])
    y_prob_list[2] = y_score_thirties

    y_score_fourties = gmm_fourties.score(X_age_test[int(x):])
    y_prob_list[3] = y_score_fourties

    y_score_fifties = gmm_fifties.score(X_age_test[int(x):])
    y_prob_list[4] = y_score_fifties

    y_score_sixties = gmm_sixties.score(X_age_test[int(x):])
    y_prob_list[5] = y_score_sixties

    # Seleccion de la puntacion mas alta y construccion del vector de etiquetas
    predichas por el clasificador para comparar con los valores de test reales
    y_prob_list_max = y_prob_list.index(max(y_prob_list))
    print('El máximo es ', max(y_prob_list), ' en la posición', y_prob_list_max)
    print('Añadiendo muestra ', x)

    if y_prob_list_max == 0:
        y_age_pred.append('teens')
        print('Muestra añadida teens')
    elif y_prob_list_max == 1:
        y_age_pred.append('twenties')
        print('Muestra añadida twenties')
    elif y_prob_list_max == 2:
        y_age_pred.append('thirties')
        print('Muestra añadida thirties')
    elif y_prob_list_max == 3:
        y_age_pred.append('fourties')
        print('Muestra añadida fourties')
    elif y_prob_list_max == 4:
        y_age_pred.append('fifties')
        print('Muestra añadida fifties')
    elif y_prob_list_max == 5:
        y_age_pred.append('sixties')
        print('Muestra añadida sixties')

    y_prob_list = [0,0,0,0,0,0]
```

Con el fragmento superior se obtiene una lista con las etiquetas predichas por el modelo para los datos de *test*. Para evaluar el rendimiento del modelo se utiliza la utilidad *Classification Report*, de igual forma que en los casos anteriores.

```
print(classification_report(y_age_test, y_age_pred))
```

```
              precision    recall  f1-score   support

     fifties       0.00      0.00      0.00       319
    fourties       0.33      0.02      0.04       334
     sixties       0.00      0.00      0.00       326
       teens       0.16      0.96      0.28       323
    thirties       0.21      0.04      0.07       358
    twenties       0.18      0.01      0.01       340

    accuracy                           0.17      2000
   macro avg       0.15      0.17      0.07      2000
weighted avg       0.15      0.17      0.07      2000
```

4.2.3.5. Resultados en el conjunto Common Voice en español

El proceso para el conjunto de datos en español es similar al mostrado en el apartado anterior, por lo que se omiten los pasos anteriores al resultado del modelo, que se muestra a continuación con la salida de la herramienta *Classification Report*.

```
print(classification_report(y_age_test, y_age_pred))
```

```
              precision    recall  f1-score   support

     fifties       0.96      0.77      0.86       897
    fourties       0.00      0.00      0.00       831
     sixties       1.00      0.84      0.92        77
       teens       0.00      0.00      0.00       107
    thirties       0.00      0.00      0.00       586
    twenties       0.42      1.00      0.59      1216

    accuracy                           0.53      3714
   macro avg       0.40      0.44      0.39      3714
weighted avg       0.39      0.53      0.42      3714
```

Capítulo 5

Resultados

5.1. Análisis de SPA comerciales

En el capítulo 3 se realizaron los análisis de seguridad y privacidad de los tres asistentes personales más utilizados del mercado. Las pruebas realizadas [Sección 3.1] han permitido obtener una visión global de las características de seguridad y privacidad que incorporan los asistentes, así como evaluar que posibilidades tiene un usuario para interactuar con ellos y como puede controlar dicha interacción.

En esta sección, el enfoque se ha centrado en evaluar los resultados obtenidos en las pruebas, comparando el comportamiento de cada asistente ante las mismas, para corroborar las hipótesis iniciales obtenidas en el estado del arte acerca de la falta de seguridad en el medio de comunicación del usuario con el asistente y los problemas de autenticación y autorización que se presentan en este ecosistema [13] [Sección 2.1].

El análisis realizado en el capítulo 3 se presenta en forma resumida en las Tablas 5.1 y 5.2, donde se recogen los resultados de las pruebas para los tres asistentes personales. En la primera tabla se muestran las pruebas realizadas acerca de la instalación e interacción con el SPA, mientras que en la segunda tabla se recogen los resultados de las pruebas de funcionalidad y seguridad en los asistentes.

En las secciones siguientes se presenta el análisis comparativo para cada una de las categorías de forma más extendida. Especialmente relevantes para el proyecto son los resultados obtenidos en las pruebas de funcionalidad y seguridad [Secciones 5.1.2.2, 5.1.2.4 y 5.1.2.5] que, como se ha expuesto al comienzo de esta sección, han permitido validar las hipótesis recogidas durante el estudio del estado del arte.

5.1.1. Pruebas de instalación e interacción

Tabla 5.1: Comparativa de características de instalación e interacción de SPA comerciales

Característica	*Google Home Mini*	*Amazon Echo Show 5*	*Apple HomePod Mini*
Instalación			
Instalación del SPA	Es necesaria una cuenta de *Google*. Los datos de *Wi-Fi*, preferencias de la cuenta de *Google* y ajustes adicionales se comparten desde el dispositivo móvil	Es necesaria una cuenta de *Amazon* para utilizar el *Echo Show 5*. Los datos de *Wi-Fi* se introducen en el dispositivo y los ajustes de *Alexa* se sincronizan desde el dispositivo móvil	Es necesaria una cuenta de *iCloud* y un dispositivo de *Apple*. Los datos de *Wi-Fi*, *Siri* y preferencias se comparten desde el *iPhone*
Instalación de nuevos dispositivos conectados	No se pueden configurar permisos para los usuarios de la casa. Cualquiera puede añadir o editar accesorios	Sólo el administrador puede añadir dispositivos y editarlos con la aplicación *Alexa*	Se permite configurar si un usuario tiene permiso para añadir o editar dispositivos
Instalación de *skills* de terceros	No hay un proceso de instalación de acciones (*skills*) de terceros. Conociendo la frase de activación puede usarse cualquier acción	El administrador puede permitir el uso de *skills* desde la aplicación *Amazon Alexa* antes de usarlas, pero un usuario cualquiera puede activar una *skill* desde el *Echo Show 5* sin confirmación, incluso puede activar de nuevo una *skill* deshabilitada	No se pueden configurar *skills* de terceros
Interacción			
Interacción con el SPA y dispositivos conectados	Se puede desactivar la reproducción de contenido multimedia en el dispositivo. No se pueden definir permisos desde *Google Home*, siendo necesario crear una familia en la aplicación externa *Family Link* y configurar filtros por dispositivo en la aplicación *Google Home*	El *Echo Show 5* dispone de controles para apagar cámara y micrófono. No hay posibilidad de distinguir entre usuarios ni de definir permisos para interacciones específicas	Se puede desactivar la invocación por voz o el control táctil. Se permite desactivar el uso de dispositivos multimedia y particularizar por usuario el control de accesorios conectados
Interacción con *skills* de terceros	No hay controles de acciones de terceros por usuario, es necesario generar un filtro de contenido para todo el grupo de usuarios de la casa	El administrador puede permitir el uso de *skills* desde la aplicación *Amazon Alexa* antes de usarlas, pero un usuario cualquiera puede activar una *skill* desde el *Echo Show 5* sin confirmación, incluso puede activar de nuevo una *skill* deshabilitada	No hay posibilidad de interactuar con *skills* de terceros

5.1.1.1. Proceso de instalación del SPA

El proceso de instalación de los asistentes personales es similar en los tres casos, como se expone en la tabla 5.1. Para poder utilizar los SPA es obligatorio disponer de una cuenta del fabricante en cuestión, que puede crearse de forma gratuita en todos los casos. Si es conveniente resaltar que mientras el *Google Home Mini* y el *Echo Show 5* pueden ser instalados con dispositivos móviles de diversos sistemas operativos, el *HomePod Mini* requiere de un dispositivos móvil de *Apple* para ser utilizado.

5.1.1.2. Instalación de nuevos dispositivos conectados

El control a la hora de instalar nuevos dispositivos difiere de forma considerable entre dispositivos. La aproximación del *HomePod Mini* es la única que combina usabilidad y seguridad, permitiendo diferenciar que usuarios en el hogar pueden instalar o editar dispositivos desde la aplicación y cuáles no. En el *Echo Show 5* no existe la posibilidad de añadir usuarios al hogar, por lo que sólo el usuario administrador puede añadir o editar accesorios. En el *Google Home Mini*, sin embargo, cuando se añade un nuevo usuario a la casa se convierte en administrador, pudiendo instalar, editar y eliminar dispositivos conectados, no habiendo posibilidad de definir permisos más restringidos a determinados usuarios.

5.1.1.3. Instalación de skills de terceros

La evaluación del proceso de instalación de *skills* de terceros en los asistentes ha proporcionado resultados que han permitido corroborar la importancia de estudios realizados sobre la seguridad de esta funcionalidad [24]. Exceptuando el *HomePod Mini*, que no cuenta con soporte de *skills* de terceros, el *Google Home Mini* y *Echo Show 5* sí permiten su uso. Si bien el proceso de instalación de *skills* de terceros en estos dos dispositivos se diferencia en la especificación, en ambos se llega al mismo resultado, un proceso de instalación de *skills* inseguro que no permite controlar quién puede añadir desarrollos de terceros al asistente. En el *Google Home Mini* no hay proceso de instalación de *skills*. Conociendo la frase de activación de la *skill* se puede activar sin controles adicionales. En el *Echo Show 5* si existe un proceso de instalación de *skills* desde la aplicación móvil, donde además se permite ver que *skills* se encuentran activadas. Pero este proceso de control pierde su sentido si se interactúa con el asistente de forma directa, ya que de esta forma no se comprueba si una *skill* ha sido activada o desactivada por el administrador. Como en el *Google Home Mini*, basta con conocer la frase de activación para utilizar la *skill*. Aún si el administrador la desactivó desde la aplicación móvil, se puede volver a activar desde el asistente simplemente invocando la *skill* de nuevo.

5.1.1.4. Interacción con el SPA y dispositivos conectados

Con estas pruebas se ha medido el nivel de control de la interacción con el SPA que es posible establecer. Como se muestra en la tabla 5.1, los tres dispositivos ofrecen enfoques diferentes. En el *HomePod Mini* se permite activar y desactivar la interacción por voz o el control táctil, además de desactivar la reproducción de contenido multimedia para ciertos usuarios y el uso de determinados dispositivos. En el *Google Home Mini* se cuenta con un botón físico para desactivar los micrófonos del dispositivo. Se puede desactivar la reproducción de contenido multimedia en el dispositivo, pero es necesario crear una familia en una aplicación externa para incorporar a los usuarios que se quiera controlar. Dicho control se configura en forma de filtros, que se aplican por grupos de usuarios (todos o miembros de la familia), no siendo posible establecer controles individuales. En el *Echo Show 5* se proporcionan controles físicos para activar y desactivar cámara y micrófonos, pero al no haber un registro de usuarios, no es posible definir permisos de interacción para usuarios específicos.

5.1.1.5. Interacción con skills de terceros

Como se ha expuesto en la Sección 5.1.1.3 y en la tabla 5.1, no es posible utilizar *skills* de terceros en el *HomePod Mini*. En el *Google Home Mini*, al no haber controles por usuarios, no se puede distinguir el uso de *skills* de forma individualizada. Aplicando filtros en el dispositivo se puede desactivar el uso de *skills* de terceros para todos los usuarios o los miembros de la casa registrados. En el *Echo Show 5* se disponen de controles para activar y desactivar *skills*, pero tal como se ha explicado en la Sección 5.1.1.3, los controles no tienen efecto si se interactúa de forma directa con el asistente, ya que no se realiza una comprobación de que usuario está intentando activar la *skill*.

5.1.2. Funcionalidad, seguridad y privacidad

Tabla 5.2: Comparativa de características de funcionalidad y privacidad y seguridad de SPA comerciales

Característica	*Google Home Mini*	*Amazon Echo Show 5*	*Apple HomePod Mini*
Funcionalidad			
Pagos y transacciones	Se pueden configurar pagos desde el asistente. Soportan un método de autenticación adicional basado en el *hardware* del dispositivo donde se haya instalado la aplicación *Google Home*	Se pueden realizar pagos con el asistente a través de *Amazon 1 Click*. Se pueden configurar métodos de confirmación adicionales a través del perfil de voz o de un código de cuatro dígitos	No se permiten los pagos o compras desde el *HomePod Mini*
Posibilidad de crear perfiles "seguros" para menores	Opción sólo disponible en *Android*. En el resto de dispositivos es necesario crear filtros de contenido de forma manual que afectan a todos los usuarios por igual, aunque opciones como pagos siguen estando activas	No hay una opción dedicada para establecer un perfil de uso seguro para menores. Es necesario desactivar y restringir ajustes en cada una de las categorías (reproducción de contenido multimedia, navegador web, pagos, *skills*, etc.) No es posible restringir el uso de accesorios conectados	No hay posibilidad de crear usuarios para menores, es necesario acceder a cada control y desactivarlo de forma manual
Privacidad y seguridad			
Respuestas que contengan información personal	Se permite desactivar las respuestas personales en los ajustes del *Google Home Mini*	No existe la posibilidad de desactivar respuestas que contengan información personal. Es necesario desactivar las funcionalidades por completo, ya que cualquier usuario puede invocarlas	Es necesario activar el reconocimiento de voz para dar respuestas con datos personales. No disponible en español
Métodos de autenticación	El dispositivo permite autenticación por voz, aunque contando con una grabación de un usuario invocando al asistente se le puede suplantar, pudiendo realizar cualquier consulta	El *Echo Show 5* cuenta con reconocimiento de voz, pero sólo se utiliza para funciones de personalización, por ejemplo en *skills*. No se utiliza para tareas de seguridad	Autenticación por reconocimiento de voz (no disponible en español)
Filtrado de voz no humana (voz de usuario pregrabada, sistemas TTS)	No se filtran mensajes procedentes de grabaciones ni de voces sintéticas	No se filtran mensajes grabados ni aquellos procedentes de voces sintéticas	Un mensaje de activación pregrabado o leído con un sistema TTS puede invocar al asistente
Interacción con el historial de conversaciones	Se incluyen opciones completas para visualizar, pausar y eliminar el historial de forma automática	En la aplicación se ofrecen opciones completas de visualización, borrado y pausado del historial de conversaciones, disponiendo de opciones de borrado automático	Se permite enviar una petición para eliminar el historial de los servidores. No se puede visualizar el historial de conversaciones

5.1.2.1. Pagos y transacciones

La posibilidad de realizar pagos desde el asistente no se encuentra disponible en el *HomePod Mini*. En el *Google Home Mini*, esta opción viene desactivada por defecto, y se puede configurar desde la aplicación móvil. Sólo el propio usuario puede activar los pagos con el asistente, ya que es necesario introducir un método de pago vinculado a su cuenta de *Google*, pero una vez configurado, cualquier usuario puede utilizarlos. Para proteger de un uso indebido, se incluye la posibilidad de autenticarse con un segundo factor en el dispositivo móvil al que esté vinculada la cuenta, utilizando para ello las capacidades del propio terminal móvil, como pueden ser el reconocimiento de huella o facial. En el *Echo Show 5* también se ofrece la posibilidad de realizar compras, para lo que es necesario activar el método de pago *Amazon 1 Click* desde la cuenta en la web de *Amazon*. Se dispone de mecanismos de autenticación basados en un código de cuatro dígitos que será requerido por el asistente o en el perfil de voz del usuario que esté realizando la compra. Si el perfil de voz se corresponde con el del usuario que ha configurado su modelo de voz en la aplicación, la compra será aceptada.

5.1.2.2. Posibilidad de crear perfiles "seguros" para menores

A fecha actual y en España, no existe posibilidad de crear perfiles para menores con uso restringido de forma sencilla y rápida en ninguno de los asistentes personales. Para restringir el uso de características en el *HomePod Mini*, es necesario ajustar las opciones una a una, habiendo casos en los que es necesario desactivar una funcionalidad para todos los usuarios del dispositivo, por ejemplo, la reproducción de contenido multimedia explícito tiene que ser desactivada para todas las peticiones que se hagan al *HomePod Mini*, no pudiendo diferenciar por usuarios. Desde la aplicación móvil se puede restringir el uso de accesorios conectados para miembros de la familia, para evitar que los menores puedan interactuar con elementos como puertas o ventanas desde su dispositivo móvil, pero nada les impide actuar sobre los dispositivos con comandos de voz, para los que no existe ningún tipo de control, más allá de desactivar por completo la activación del asistente. En el *Google Home Mini*, la opción de incluir menores en el grupo de usuarios del asistente sólo está disponible en *Android*. En el resto de dispositivo es necesario establecer filtros, que como se ha visto anteriormente, afectan por igual a todos los usuarios del asistente. Con los filtros se puede restringir la reproducción de contenido multimedia, las llamadas y el uso de *skills* de terceros. Sin embargo, como se ha expuesto en la tabla 5.2, los pagos siguen estando activos en el asistente a no ser que se desactiven para todos los usuarios. El *Echo Show 5* tampoco dispone de la posibilidad de establecer perfiles restringidos para menores, por lo que hay que ajustar las distintas configuraciones de forma manual. Las restricciones se configuran desde el propio dispositivo, siendo necesario introducir la contraseña de la cuenta de *Amazon* vinculada cada vez que se quieran realizar cambios. Es posible restringir el uso del navegador web, la reproducción de contenido multimedia o los pagos. En cuanto a los accesorios conectados, como en el resto de dispositivos, no es posible restringir su uso.

5.1.2.3. Respuestas que contengan información personal

El tratamiento de las respuestas que contengan información personal puede dividirse en dos grupos, formados uno por el *HomePod Mini* y el *Google Home Mini* y otro por el *Echo Show 5*. En el *HomePod Mini* y en el *Google Home Mini* las respuestas personales están relacionadas con el ajuste de reconocimiento de voz que se tenga seleccionado. Si el usuario tiene el perfil de voz guardado y el reconocimiento de voz está activo, cuando el asistente sea preguntado por información como eventos del calendario, mensajes o recordatorios, entre otros ejemplos, se comprobará que usuario inició la consulta, para ofrecer una respuesta acorde a los datos de esa persona. Si otro usuario realiza una pregunta sobre información personal, recibirá una respuesta con su propia información, en caso de que su perfil de voz esté almacenado, o se le denegará el acceso a la información de la respuesta si se trata de un usuario desconocido para el asistente. Al contrario que en los dos dispositivos descritos, en el *Echo Show 5* no se dispone de ningún método de control de las respuestas que contengan información personal. La única opción para evitar que usuarios ajenos puedan acceder a dicha información es no usar el servicio que pueda generarla.

5.1.2.4. Métodos de autenticación

La autenticación en los tres asistentes analizados se realiza a través del reconocimiento de voz. En el *HomePod Mini* la autenticación por reconocimiento de voz no está disponible en español, mientras que en el *Google Home Mini* y en el *Echo Show 5* si se ofrece esta posibilidad. En condiciones de uso normales, los dispositivos son capaces de distinguir entre diferentes voces y reconocer a usuarios, denegando el acceso a información personal entre usuarios de forma satisfactoria. La autenticación no se utiliza generalmente como una medida de seguridad, sino como una funcionalidad accesoria para personalizar la experiencia con el

asistente. Esto es así debido a la inseguridad de la función de reconocimiento de voz, como se ha comprobado de forma explícita en el análisis y como se indica en la documentación de *Google* [21]. Como se expone en el apartado siguiente, todos los asistentes son susceptibles a ser activados con mensajes grabados o procedentes de sistemas *Text To Speech*. Este problema, combinado con un comportamiento de la autenticación deficiente, que sólo está activa durante la frase de activación y no durante el mensaje completo, provoca una situación en la que cualquier persona con una grabación del usuario legítimo pronunciando el mensaje de activación pueda suplantarle, sin necesidad de tener el mensaje completo grabado. El proceso de suplantación de un usuario descrito se indica de forma resumida en los diagramas de las figuras 5.1 y 5.2.

Figura 5.1: Captura del mensaje de activación de un usuario legítimo

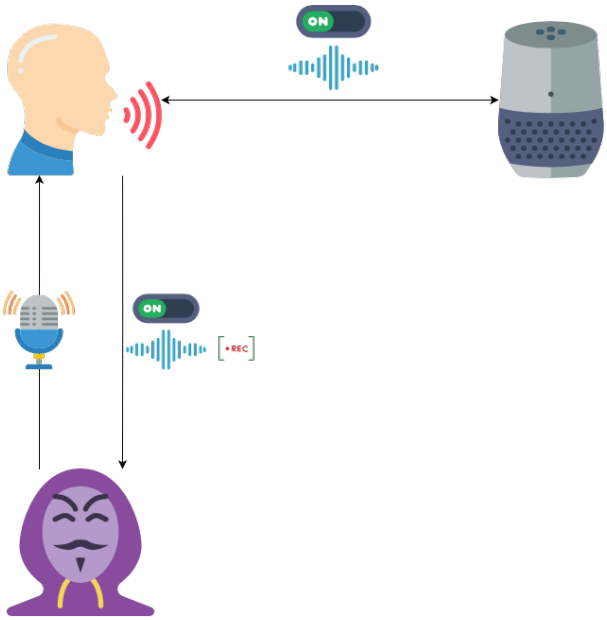
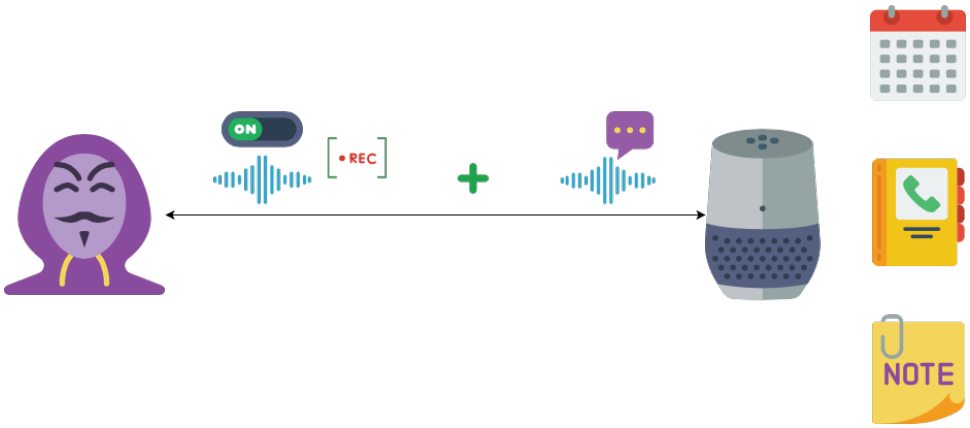


Figura 5.2: Ataque de suplantación con grabación del mensaje de activación



El procedimiento de suplantación descrito ha sido probado en los asistentes que cuentan con reconocimiento de voz en español (*Google Home Mini* y *Echo Show 5*), resultando satisfactorio en ambos.

5.1.2.5. Filtrado de voz no humana

Se ha comprobado que los tres SPA analizados son susceptibles a ser activados mediante voces de origen artificial, como grabaciones y sistemas *Text To Speech*. No implementan ningún tipo de medida de protección frente a voces sintéticas más allá de apagar los micrófonos de los dispositivos cuando no se estén controlando físicamente. Este hallazgo, que por si sólo no puede parecer que presenta un problema grave, combinado con un proceso de autenticación deficiente, como se ha detallado en el apartado anterior, acarrea la posibilidad de que un actor malicioso que cuente con una grabación del mensaje de activación del usuario pueda realizar cualquier consulta al asistente, evadiendo los controles de reconocimiento de voz y pudiendo extraer información personal del dispositivo, interactuar con accesorios conectados en la casa, etc.

5.1.2.6. Interacción con el historial de conversaciones

Las opciones de interacción con el historial de conversaciones son diferentes en cada uno de los dispositivos. El *HomePod Mini* es el que ofrece menos posibilidades, pudiendo únicamente enviar una petición para eliminar las grabaciones de las conversaciones almacenadas en los servidores de *Apple*. En el *Google Home Mini* se ofrece una sección completa de ajustes de privacidad para visualizar y eliminar las grabaciones de voz, así como configurar un borrado automático de las grabaciones y pausar el almacenamiento de las mismas. Las opciones que *Amazon* incorpora en el *Echo Show 5* son similares a las de *Google*, ofreciendo un conjunto completo de opciones para revisar las conversaciones con el asistente, eliminarlas y configurar un borrado automático de los datos.

5.2. Resultados del clasificador de voces

En el capítulo 4 se ha descrito el proceso completo de creación y validación de varios modelos de clasificación de voz que permiten distinguir entre múltiples grupos de edad. Seleccionando conjuntos de datos en inglés y español, así como diferentes herramientas de clasificación se han conseguido resultados relevantes para comenzar una línea de trabajo que podrá ser continuada en el futuro a partir de los avances conseguidos en el estudio, que permitirán llegar a conseguir una herramienta comercial robusta y fiable que realice una distinción de usuarios menores y adultos en los asistentes personales comerciales de forma automática.

El objetivo de esta sección es exponer de forma conjunta los resultados obtenidos para cada uno de los modelos de clasificación con el fin de analizarlos y extraer conclusiones de las técnicas aplicadas y el rendimiento obtenido. Los resultados de los modelos de clasificación basados en SVM y *Random Forest* se encuentran resumidos en las tablas 5.3 y 5.4 respectivamente.

5.2.1. Clasificación con Support Vector Machines

En la tabla 5.3 y en las figuras 5.3, 5.4 y 5.5 se muestran los resultados obtenidos para los tres clasificadores de voz desarrollados con modelos basados en *Support Vector Machines*. Como se puede observar, los resultados obtenidos difieren considerablemente entre idiomas, siendo la clasificación con el conjunto de datos en español mucho más precisa que la clasificación con el conjunto en inglés, tanto en la versión de tamaño similar al conjunto en español como en su versión extendida, que si bien mejora el resultado, queda muy por debajo del rendimiento alcanzado por el clasificador en español. Utilizando como métrica principal la precisión ponderada con el soporte de cada una de las clases, el clasificador SVM en español obtiene una precisión del 94 %, mientras que el clasificador en inglés alcanza una precisión del 59 % y el clasificador en ingles con la versión extendida del *dataset* consigue una precisión del 65 %. Siendo el proceso de extracción de características exactamente igual para los tres *datasets*, estos resultados pueden deberse a que el conjunto de datos en inglés contenga más tipos de voces diferentes, así como grabaciones de calidades distintas que puedan degradar el rendimiento haciendo la tarea clasificación más difícil. Además, como cabía esperar, la precisión de la clasificación no es igual para todos los grupos de edad. En los tres conjuntos de datos analizados se observa el mismo comportamiento, en los grupos de edad más bajos (`teens`, `twenties` y `thirties`), la precisión es menor que en el resto de grupos [Tabla 5.3]. En las figuras 5.3, 5.4 y 5.5 se observa esta situación, donde los grupos de edad más bajos comparten una gran proporción de las muestras erróneas con aquellos adyacentes (la mayor parte de errores de clasificación del grupo `twenties` provienen del grupo `thirties` y viceversa).

Tabla 5.3: Precisión por grupo de edad con modelos SVM

Grupos de edad	*Common Voice* Español	*Common Voice* Inglés	*Common Voice* Extendido
Teens	90 %	48 %	53 %
Twenties	92 %	45 %	52 %
Thirties	91 %	45 %	54 %
Fourties	97 %	80 %	82 %
Fifties	98 %	51 %	65 %
Sixties	98 %	83 %	87 %
Precisión ponderada	**94 %**	**59 %**	**65 %**

Figura 5.3: Matriz de confusión del clasificador SVM. *Common Voice* en español

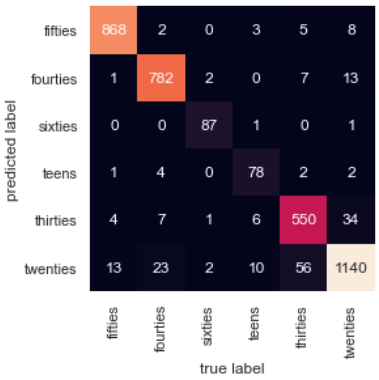


Figura 5.4: Matriz de confusión del clasificador SVM. *Common Voice* en inglés

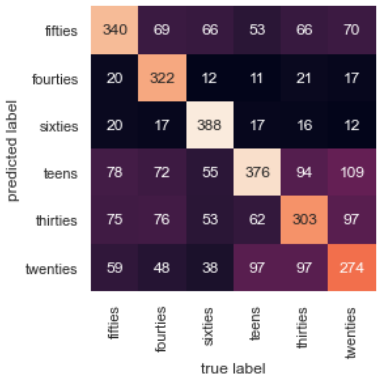
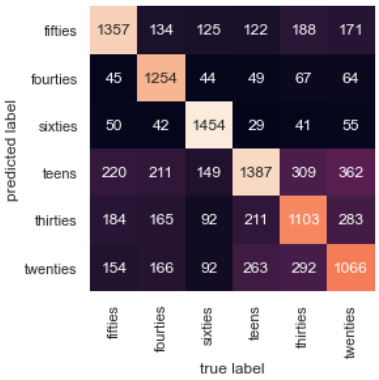


Figura 5.5: Matriz de confusión del clasificador SVM. *Common Voice* en inglés extendido



5.2.2. Clasificación con Random Forest

Los resultados de los clasificadores de voz desarrollados con *Random Forest* se muestran en la tabla 5.4 y en las figuras 5.6, 5.7 y 5.8. Se observa un rendimiento similar al alcanzado con SVM, con una precisión con el conjunto en español del 94 %, una precisión del 57 % con el *dataset* en inglés y del 63 % con el conjunto en inglés extendido. Si bien la precisión global es algo más reducida con los modelos de *Random Forest*, cabe destacar que se obtiene una mayor precisión para el grupo de edad `teens` en todos los modelos, lo que es especialmente relevante para el objetivo de distinción entre menores y adultos. Como en el caso de SVM, los grupos de edad `twenties` y `thirties` son los más difíciles de clasificar, por la proximidad de las voces en ambos grupos. En las matrices de confusión de las figuras 5.6, 5.7 y 5.8 se recoge la tendencia descrita, localizándose la mayor parte de errores en los grupos de edad más bajos entre aquellos contiguos.

Tabla 5.4: Precisión por grupo de edad con modelos *Random Forest*

Grupos de edad	*Common Voice* Español	*Common Voice* Inglés	*Common Voice* Extendido
Teens	98 %	60 %	63 %
Twenties	88 %	44 %	51 %
Thirties	93 %	42 %	50 %
Fourties	98 %	77 %	83 %
Fifties	99 %	47 %	56 %
Sixties	100 %	71 %	76 %
Precisión ponderada	**94 %**	**57 %**	**63 %**

Figura 5.6: Matriz de confusión del clasificador *Random Forest*. *Common Voice* en español

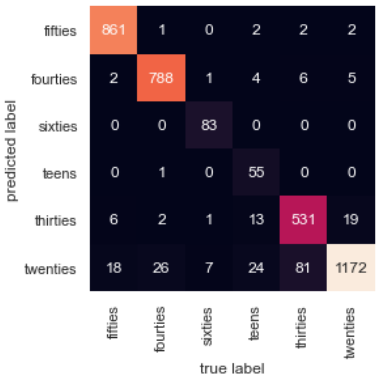


Figura 5.7: Matriz de confusión del clasificador *Random Forest*. *Common Voice* en inglés

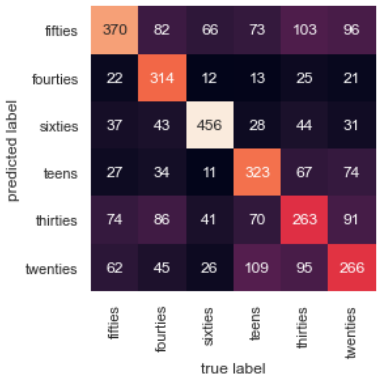
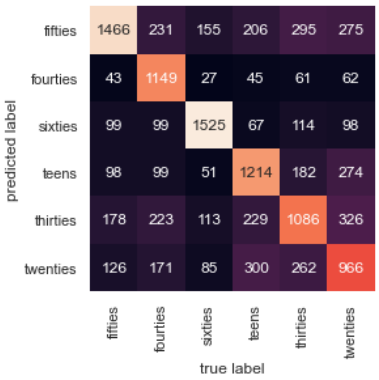


Figura 5.8: Matriz de confusión del clasificador *Random Forest*. *Common Voice* en inglés extendido



5.2.3. Clasificación con Gaussian Mixture Models

Como se ha podido observar en la Sección 4.2.3 en el capítulo 4, los resultados de la clasificación con modelos estadísticos no son satisfactorios, ofreciendo un resultado errático y pobre para todos los grupos de edad. Estos resultados pueden deberse a la gran cantidad de voces que pueden darse dentro de cada uno de los grupos de edad y las diferencias en la calidad de las grabaciones. En los estudios analizados de clasificación de voz que hacen uso de modelos GMM, los *datasets* cuentan con ejemplos grabados en las mismas condiciones técnicas y ambientales, además de tener un número de voces relativamente limitado, lo que facilita la creación de modelos de voz más fiables. En el conjunto de datos utilizado, al tratarse de una plataforma abierta y colaborativa, cada usuario ha grabado con su propio equipo la voz que ha donado al conjunto de datosz, y además se cuenta con un número de voces diferentes mucho más elevado.

Capítulo 6

Conclusiones y trabajos futuros

El trabajo se ha realizado dentro del marco del proyecto europeo RAYUELA, contribuyendo específicamente en el paquete de trabajo *Technology assessment and IT threat landscape* [18], cuyo objetivo es identificar y evaluar los riesgos en los dispositivos conectados utilizados por menores, donde se engloban los *Smart Personal Assistants* analizados en este proyecto. Los avances realizados en el paquete de trabajo del proyecto RAYUELA y los resultados del estudio han dado lugar a un artículo publicado en las Jornadas Nacionales de Investigación en Ciberseguridad, celebradas en Junio de 2021, donde se analizan desde el punto de vista de la seguridad y privacidad los dispositivos IoT utilizados por jóvenes [19].

Adicionalmente se preparará un artículo centrado en el análisis de los asistentes personales inteligentes comerciales llevado a cabo en el proyecto, que será enviado a conferencias como *ACM CCS*, *IEEE S&P*, *ACM CHI* y *IEEE WFIoT* o a revistas como *IEEE IoT Journal* y *ACM Transactions on Internet of Things*.

6.1. Conclusiones del análisis de dispositivos comerciales

El análisis de dispositivos comerciales ha permitido comprobar las hipótesis iniciales acerca de la inseguridad en los asistentes personales que se habían identificado en las referencias consultadas al comienzo del proyecto. Especialmente relevantes son los hallazgos en lo que respecta a los sistemas de autenticación de usuarios que los asistentes incorporan, resultando deficientes sus medidas de protección frente a intentos de suplantación con una grabación de un usuario legítimo.

Además, la protección de usuarios menores a través de ajustes sencillos y rápidos es prácticamente inexistente en las condiciones analizadas (versiones de los dispositivos, versiones de las aplicaciones y dispositivo móvil al que se han enlazado). En todos los casos es necesario recorrer el conjunto completo de ajustes de los dispositivos, incluso teniendo que cambiar entre menús de configuración diferentes, para desactivar todas las características que pueden resultar inseguras para un usuario menor sin supervisión. Se ha podido comprobar como en muchos casos, las configuraciones restringidas que se hagan de los dispositivos afectan por igual a todos los usuarios, dejando características inutilizadas para todos los miembros del hogar, lo que no resulta práctico y no favorece que los usuarios establezcan controles de uso. Por tanto, en este proyecto se han identificado dos líneas de trabajo complementarias que ayudarán a proteger a los usuarios menores en los SPA:

- Es necesario definir diferentes perfiles de seguridad preconfigurados con políticas de acceso que permitan controlar de forma sencilla pero precisa como los usuarios se relacionan con los dispositivos, tal como se hace en otras plataformas de contenido multimedia, como *Netflix* o *HBO*. En el ecosistema de los SPA la creación de perfiles de seguridad para menores es más compleja, ya que las funcionalidades de los dispositivos, y por tanto las necesidades de control, son más variadas y complejas que en una aplicación de reproducción multimedia, por lo que es necesario analizar exhaustivamente las características de los SPA para identificar y evaluar los riesgos que su uso por parte de un menor puede producir.

- Ante la necesidad de recurrir a las muestras de voz que el asistente captura en el momento de la invocación por parte de un usuario para autenticarle, es preciso desarrollar un sistema robusto que permita distinguir a usuarios menores y mayores de edad cada vez que se envía una consulta al asistente, para aplicar los perfiles de seguridad de forma automática. Con esto se favorecería el uso de las opciones de seguridad incluidas, ya que se aplicarían de forma transparente para los usuarios, tal como se aplican los mecanismos de autenticación en los *smartphones* actuales.

Es por ello que la línea de investigación iniciada en el proyecto acerca de la distinción automática entre usuarios menores y adultos resulta de especial relevancia en el ecosistema de los SPA, ya que permitirá que los usuarios puedan disfrutar de la funcionalidad completa de los asistentes, mientras que se protege a los usuarios menores para que hagan un uso adecuado de los dispositivos, sin necesidad de tener que ajustar manualmente las configuraciones y permisos de usuarios, ya que cuando se detecte la voz de un menor con la herramienta de clasificación, se aplicarán los permisos de uso restringidos de forma automática.

6.2. Conclusiones del clasificador de voz

Se han utilizado tres técnicas diferentes para desarrollar clasificadores que distingan a usuarios por edad. Los modelos basados en *Support Vector Machines* y *Random Forest* han ofrecido los mejores resultados. También se ha comprobado como para el conjunto de datos en español se obtienen rendimientos más altos, cercanos al 95 % de precisión con ambas herramientas. Para los conjuntos en inglés, el rendimiento disminuye considerablemente hasta situarse entorno al 57-65 %.

Con este estudio no se ha buscado desarrollar una herramienta infalible que pudiera ser usada en el ámbito comercial de forma directa, sino establecer un punto de partida en la línea de investigación de la clasificación de voz para que en futuros trabajos se continúen depurando los modelos, haciendo uso de nuevas técnicas de Inteligencia Artificial, conjuntos de datos alternativos y extrayendo características diferentes que permitan aumentar el rendimiento.

Los resultados alcanzados demuestran que la clasificación de voces por grupos de edad es posible, y los rendimientos obtenidos con conjuntos de datos y herramientas abiertas posibilitan un trabajo de investigación fructífero en este área.

6.3. Trabajos futuros

Durante el desarrollo del proyecto se han identificado diferentes líneas de trabajo futuras, que provienen de puntos de mejora de las soluciones aportadas en el proyecto y de nuevas áreas que se han descubierto durante la realización del mismo. En la lista siguiente se exponen los trabajos futuros identificados más relevantes, tanto para la investigación acerca de *Smart Personal Assistants* como para la clasificación de voz:

- Investigación en *Smart Personal Assistants*
    - Sistema de autenticación por voz continuo. El ataque de suplantación presentado en el proyecto demuestra que los asistentes personales comerciales no realizan la autenticación de los mensajes completos que reciben. Únicamente se comprueba la identidad del usuario en el momento de la activación, lo que permite a un actor malicioso con una grabación poder suplantar a un usuario y realizar cualquier consulta al SPA. Con un sistema de autenticación continuo, el mensaje completo que recibe el asistente debe ser autenticado, así como las posteriores interacciones en la conversación, para dificultar este tipo de suplantaciones. Para eliminar la posibilidad de que se realice este tipo de suplantaciones, además de este sistema, es necesario una capa de procesamiento de los mensajes adicional, que permita identificar si la voz del mensaje procede de una fuente artificial o es una voz humana.
    - Desarrollo de un marco de evaluación de funcionalidades para usuarios menores de edad. La detección de usuarios menores en los asistentes personales debe ir acompañada de un conjunto de ajustes de seguridad robustos, que deben estudiarse y presentarse en un modelo estandarizado que permita definir que características de los asistentes representan un riesgo para los menores y que configuración debe tener cuando estén siendo utilizadas por un menor.
- Clasificación de voz
    - Pruebas con *datasets* especializados y técnicas de filtrado de muestras de voz. La variabilidad en las condiciones de grabación de las muestras de audio en los conjuntos de datos y el amplio abanico de edades que se incluyen en la categoría de menores del conjunto *Common Voice*, pueden ser una causa de disminución del rendimiento de los modelos, como se ha visto en el proyecto. Para obtener un rendimiento óptimo, sería necesario contar con *datasets* que contengan muestras especializadas en niños, donde exista un etiquetado riguroso de las edades que permita identificar de forma más precisa a los menores [55]. También hay que tener en cuenta que un clasificador implementado en un SPA no va a funcionar en condiciones ideales, ya que existirá ruido en el ambiente donde esté implementado o la distancia a la que los usuarios hablen no será siempre la misma, por lo que es importante estudiar técnicas de filtrado del audio para asegurar que la muestra que procesa el asistente sea lo más limpia posible con el fin de evitar confusiones o potenciales comportamientos maliciosos.
    - Investigación de ataques con *adversarial AI* a los modelos de clasificación. Las técnicas de *adversarial AI* han tomado importancia en los últimos tiempos, representando una amenaza para los sistemas de Inteligencia Artificial, tanto en entrenamiento como en producción. Analizar como este tipo de técnicas afectan al rendimiento de los modelos, y estudiar posibles líneas de defensa frente a estos ataques preparará a los modelos ante situaciones venideras en las que los ataques por *adversarial AI* sean cada vez más frecuentes.
    - Desarrollo de un sistema de detección de voces sintéticas. Los descubrimientos realizados durante el desarrollo de las pruebas en los asistentes personales comerciales analizados revelan una necesidad importante de un sistema capaz de diferenciar entre voces procedentes de fuentes sintéticas y voces humanas. Un sistema de este tipo, junto con la aplicación de un esquema de autenticación por voz continuo, evitará ataques de suplantación como el presentado en el proyecto.
    - Aplicación de técnicas de visualización a las muestras de audio y estudio de técnicas de *Deep Learning* para clasificación de voz por edades. El rendimiento de las técnicas de *Deep Learning* amplía las posibilidades a la hora de configurar sistemas de Inteligencia Artificial, posibilitando con el uso de redes neuronales convolucionales la extracción automática de características de las muestras de voz, por lo que se puede estudiar su aplicabilidad en el ámbito de la clasificación de voz por edades o la detección de voces sintéticas. Así mismo, el desarrollo de técnicas de visualización avanzadas sobre los datos puede ayudar a alcanzar un mayor conocimiento de su estructura, pudiendo aumentar el rendimiento de los modelos que se desarrollen.

Bibliografía

[1] S. Ruan, J. O. Wobbrock, K. Liou, A. Ng, and J. A. Landay, "Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, p. 1–23, Enero 2018.

[2] Statista, "Number of digital voice assistants in use worldwide from 2019 to 2024." https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/, Abril 2020. Último acceso: 9 de Marzo de 2021.

[3] Amazon, "Alexa features." https://www.amazon.com/b?ie=UTF8&node=21576558011, 2021. Último acceso: 7 de Marzo de 2021.

[4] Statista, "Smart speaker use case frequency in the united states as of january 2020." https://www.statista.com/statistics/994696/united-states-smart-speaker-use-case-frequency/, 2020. Último acceso: 9 de Marzo de 2021.

[5] S. Liu, "Smart home in the united states - statistics & facts." https://www.statista.com/topics/6201/smart-home-in-the-united-states/, Enero 2021. Último acceso: 7 de Marzo de 2021.

[6] Statista, "Smart home penetration rate forecast for selected countries 2019." https://www.statista.com/forecasts/483757/penetration-rate-of-smart-homes-for-selected-countries, 2019. Último acceso: 7 de Marzo de 2021.

[7] Statista, "Smart home devices unit shipments in the united states from 2016 to 2020." https://www.statista.com/statistics/798370/us-smart-home-units/, Enero 2020. Último acceso: 7 de Marzo de 2021.

[8] Nest, "Nest learning thermostat." https://store.google.com/us/product/nest_learning_thermostat_3rd_gen?hl=es-ES, 2021. Último acceso: 7 de Marzo de 2021.

[9] Philips, "Hue." https://www.philips-hue.com/en-us, 2021. Último acceso: 7 de Marzo de 2021.

[10] Statista, "Total number of brands compatible with amazon alexa from january 2018 to july 2020." https://www.statista.com/statistics/912903/amazon-alexa-brand-use-growth/, Enero 2021. Último acceso: 7 de Marzo de 2021.

[11] Statista, "Total number of smart home devices that are compatible with amazon's alexa as of july 2020." https://www.statista.com/statistics/912893/amazon-alexa-smart-home-compatible/, Julio 2020. Último acceso: 9 de Marzo de 2021.

[12] Statista, "Number of smart home devices supported by google assistant worldwide from january 2018 to may 2019." https://www.statista.com/statistics/933532/worldwide-google-assistant-device-support/, Mayo 2019. Último acceso: 9 de Marzo de 2021.

[13] J. S. Edu, J. M. Such, and G. Suarez-Tangil, "Smart home personal assistants: A security and privacy review," *CoRR*, vol. abs/1903.05593, 2019.

[14] Alexa, "Alexa internet privacy notice." https://www.alexa.com/help/privacy, Julio 2020. Último acceso: 8 de Marzo de 2021.

[15] Mozilla Foundation, "*privacy not included." https://foundation.mozilla.org/en/privacynotincluded/amazon-echo-dot/, Febrero 2020. Último acceso: 8 de Marzo de 2021.

[16] Statista, "Smart speaker with intelligent personal assistant market share in 2018 and 2019, by platform." https://www.statista.com/statistics/1005558/worldwide-smart-speaker-market-share/, Mayo 2019. Último acceso: 9 de Marzo de 2021.

[17] Statista, "Digital voice assistant installed base worldwide in 2019, by brand." https://www.statista.com/statistics/967412/worldwide-digital-voice-assistant-installed-base-brand/, Mayo 2020. Último acceso: 21 de Junio de 2021.

[18] RAYUELA, "About rayuela." https://www.rayuela-h2020.eu/about-us/, 2021. Último acceso: 10 de Julio de 2021.

[19] S. Solera-Cotanilla, M. Vega-Barbas, M. Á.-C. Fernández-Corredor, C. V. Amores, J. F. de la Fuente, and G. L. López, "Análisis de seguridad y privacidad en dispositivos de la internet de las cosas usados por jóvenes," in *Investigación en Ciberseguridad*, Ediciones de la Universidad de Castilla-La Mancha, 2021.

[20] X. Yuan, Y. Chen, A. Wang, K. Chen, S. Zhang, H. Huang, and I. M. Molloy, "All your alexa are belong to us: A remote voice control attack against echo," in *2018 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, 2018.

[21] Google, "Link your voice to your devices with voice match." https://support.google.com/assistant/answer/9071681#vm_pr, Marzo 2021. Último acceso: 16 de Marzo de 2021.

[22] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7, 2018.

[23] J. B. Li, S. Qu, X. Li, J. Szurley, J. Z. Kolter, and F. Metze, "Adversarial music: Real world audio adversary against wake-word detection system," 2019.

[24] N. Zhang, X. Mi, X. Feng, X. Wang, Y. Tian, and F. Qian, "Dangerous skills: Understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems," in *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 1381–1396, 2019.

[25] Amazon, "Alexa accesibility - how to change your wake word." https://www.amazon.com/-/es/b?ie=UTF8&node=21341305011, 2021. Último acceso: 1 de Junio de 2021.

[26] Google, "Nest hub max." https://store.google.com/us/product/google_nest_hub_max?hl=en-US, Junio 2021. Último acceso: 4 de Junio de 2021.

[27] Google, "Face match." https://support.google.com/googlenest/answer/9320885/face-match-on-google-nest-hub-max-android, Junio 2021. Último acceso: 4 de Junio de 2021.

[28] J. Liu, X. Kong, F. Xia, X. Bai, L. Wang, Q. Qing, and I. Lee, "Artificial intelligence in the 21st century," *IEEE Access*, vol. 6, pp. 34403–34421, Marzo 2018.

[29] H. Zhao and P. Wang, "A short review of age and gender recognition based on speech," in *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, (Washington, DC, USA), pp. 183–185, IEEE, 2019.

[30] S. Safavi, *Speaker Characterization using Adult and Children's Speech*. PhD thesis, University of Birmingham. School of Electronic, Electrical and Systems Engineering., 2015.

[31] D. Naik, "Pole-filtered cepstral mean subtraction," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, (Detroit, MI, USA), pp. 157–160 vol.1, IEEE, 1995.

[32] A. Garcia and R. Mammone, "Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, vol. 1, (Phoenix, AZ, USA), pp. 325–328 vol.1, IEEE, 1999.

[33] Scikit Learn, "Plot different svm classifiers in the iris dataset." https://scikit-learn.org/stable/auto_examples/svm/plot_iris_svc.html, 2021. Último acceso: 27 de Junio de 2021.

[34] Scikit Learn, "Support vector machines." https://scikit-learn.org/stable/modules/svm.html, 2021. Último acceso: 27 de Junio de 2021.

[35] C. M. Bishop, *Pattern Recognition and Machine Learning*, ch. 7. Sparse Kernel Machines, pp. 338–339. Springer Science+Business Media, LLC, 2006.

[36] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 4211–4215, 2020.

[37] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, (New York, NY, USA), p. 1459–1462, Association for Computing Machinery, 2010.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[39] Statista, "Global smartphone market share from 4th quarter 2009 to 1st quarter 2021." https://www.statista.com/statistics/271496/global-market-share-held-by-smartphone-vendors-since-4th-quarter-2009/, Junio 2020. Último acceso: 21 de Junio de 2021.

[40] Statista, "Tablet shipments market share by vendor worldwide from 2nd quarter 2011 to 1st quarter 2021." https://www.statista.com/statistics/276635/market-share-held-by-tablet-vendors/, Mayo 2021. Último acceso: 21 de Junio de 2021.

[41] Apple, "Homepod mini tech specs." https://www.apple.com/homepod-mini/specs/, Junio 2021. Último acceso: 25 de Junio de 2021.

[42] Google, "Compare smart speakers and displays." https://store.google.com/us/magazine/compare_nest_speakers_displays?hl=en-US, Marzo 2021. Último acceso: 11 de Marzo de 2021.

[43] Google, "Integrate with google assistant." https://developers.google.com/assistant, Marzo 2021. Último acceso: 14 de Marzo de 2021.

[44] Google, "Google nest or home device and your child's google account." https://support.google.com/googlenest/answer/9039704?hl=en, 2021. Último acceso: 28 de Junio de 2021.

[45] Statista, "Total number of amazon alexa skills from january 2019 to october 2020, by country." https://www.statista.com/statistics/1189221/amazon-alexa-total-skills/, Octubre 2020. Último acceso: 14 de Marzo de 2021.

[46] Amazon, "Amazon echo & alexa devices." https://www.amazon.com/smart-home-devices/b?ie=UTF8&node=9818047011, Marzo 2021. Último acceso: 14 de Marzo de 2021.

[47] Amazon, "Echo show 10." https://www.amazon.com/dp/B07VHZ41L8/ref=sxts, Marzo 2021. Último acceso: 15 de Marzo de 2021.

[48] Amazon, "Safety information for using smart home devices with alexa." https://www.amazon.com/gp/help/customer/display.html?nodeId=201819970, 2021. Último acceso: 29 de Junio de 2021.

[49] Amazon, "Buy now ordering." https://www.amazon.com/gp/help/customer/display.html?nodeId=201889620, 2021. Último acceso: 1 de Julio de 2021.

[50] Scikit Learn, "Standard scaler." https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html, 2021. Último acceso: 5 de Julio de 2021.

[51] Scikit Learn, "Classification report." https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html, 2021. Último acceso: 5 de Julio de 2021.

[52] Scikit Learn, "Classification report." https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees, 2021. Último acceso: 6 de Julio de 2021.

[53] Scikit Learn, "Gmm." https://scikit-learn.org/stable/modules/mixture.html#gmm, 2021. Último acceso: 6 de Julio de 2021.

[54] Scikit Learn, "Gaussian mixture model selection." https://scikit-learn.org/stable/auto_examples/mixture/plot_gmm_selection.html#sphx-glr-auto-examples-mixture-plot-gmm-selection-py, 2021. Último acceso: 6 de Julio de 2021.

[55] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The ogi kids' speech corpus and recognizers," in *Sixth International Conference on Spoken Language Processing*, 2000.

[56] ONU, "Objetivos de desarrollo sostenible." https://www.un.org/sustainabledevelopment/es/, 2021. Último acceso: 11 de Julio de 2021.

[57] ONU, "Industria, innovación e infraestructuras." https://www.un.org/sustainabledevelopment/es/infrastructure/, 2021. Último acceso: 11 de Julio de 2021.

[58] ONU, "Educación de calidad." https://www.un.org/sustainabledevelopment/es/infrastructure/, 2021. Último acceso: 11 de Julio de 2021.

Objetivos de Desarrollo Sostenible

Los Objetivos de Desarrollo Sostenible (ODS) [56] [Figura 1] se definieron como parte de la Agenda 2030 de las Naciones Unidas para el Desarrollo Sostenible, cuyo propósito es alcanzar un desarrollo responsable con los recursos naturales que no comprometa el progreso de las generaciones futuras. Con los ODS se proponen diecisiete líneas de acción para alcanzar tres pilares básicos: crecimiento económico, inclusión social y protección del medio ambiente.

Figura 1: Objetivos de Desarrollo Sostenible

El proyecto se alinea con el objetivo número nueve, *Industria, Innovación e Infraestructuras* [57], cuyo propósito es construir una industria sostenible, duradera y fomentar la innovación, como medios de crecimiento económico y de mejora de la calidad de vida de las personas. De forma más específica, el trabajo contribuye con la meta 9.5, aumentar la investigación científica y mejorar la capacidad tecnológica de los países. Se ha hecho un esfuerzo por utilizar herramientas de desarrollo abiertas y gratuitas para llevar a cabo el proyecto, fomentando la investigación en el área de desarrollo del trabajo, abriendo nuevas líneas de investigación que permitan mejorar la seguridad de los dispositivos analizados, contribuyendo así a una mejora tecnológica que afectará positivamente a la vida de las personas.

El proyecto contribuye además al desarrollo del objetivo número cuatro, *Educación de calidad* [58], que persigue garantizar una educación inclusiva, equitativa y de calidad, promoviendo oportunidades de aprendizaje para todas las personas. Gracias al potencial que los *Smart Personal Assistants* presentan para ser utilizados como herramientas educativas, como potenciales plataformas de acceso a información de calidad acerca de un amplio abanico de temáticas, el desarrollo propuesto en este trabajo contribuye a reforzar la seguridad de estos dispositivos, convirtiéndolos en plataformas de aprendizaje seguras y accesibles para los niños.

Presupuesto económico

El presupuesto del Trabajo Fin de Máster se indica en la Tabla 1.

Tabla 1: Presupuesto económico del Trabajo Fin de Máster

Costes		Horas	Precio/Hora	Total
Coste de mano de obra (Costes directos)		350	35 €	**12250 €**
Coste de recursos materiales (Costes directos)	Precio de compra	Uso en meses	Amortización en años	Total
Ordenador personal	2699 €	7	5	314,88 €
Total				**314,88 €**
Gastos generales (Costes indirectos)	15 % sobre Costes Directos			**1884,73 €**
Beneficio industrial	6 % sobre (Costes Directos + Costes Indirectos)			**866,98 €**
Material fungible				
HomePod Mini				99 €
Google Home Mini				59 €
Echo Show 5				64,99 €
Subtotal presupuesto				**15539,58 €**
IVA aplicable	21 %			3263,31
Total presupuesto				**18802,89**