Análisis de seguridad y privacidad de Asistentes Personales con voces reales y voces sintéticas

Clara Palacios-Castrillo, Rafael Palacios, Roberto Gesteira-Miñarro, Gregorio López
Instituto de Investigación Tecnológica, Universidad Pontificia Comillas, Madrid, SPAIN

Clara.Palacios.Castrillo@alu.comillas.edu, Rafael.Palacios@iit.comillas.edu, rgesteira@comillas.edu, gllopez@comillas.edu

Resumen- **En este artículo se muestra el comportamiento de varios asistentes personales (Smart Personal Assistants, SPAs) en diversos aspectos relativos a la seguridad y a la privacidad. Los asistentes personales se pueden entrenar con la voz del propietario, y los rasgos de la voz actúan como contraseña de acceso. Sin embargo, se ha querido analizar qué información llegan a desvelar diferentes asistentes personales sin verificar la voz del propietario y qué riesgos reales existen de suplantar la voz del propietario. Para ello se ha definido un protocolo de ensayos que incluye comandos para solicitar información genérica, solicitudes de información personal, y solicitudes más delicadas como realizar llamadas o realizar compras. Para ensayar distintos tipos de voz participaron varias personas, se hicieron grabaciones de voz y se utilizaron herramientas de generación de voz sintética. Con objeto de engañar a los asistentes personales, se han realizado ensayos con diversos sistemas de IA generativa para crear modelos de voz basados en el usuario registrado en los asistentes, de esta manera se han podido generar sintéticamente comandos con los rasgos de voz de dicha persona. En este estudio se ha trabajado con los asistentes Apple HomePod, Amazon Alexa y Google Home, que son los principales dispositivos del mercado. Se ha podido comprobar qué tipo de información comunica cada sistema sin realizar la validación al usuario y se ha podido comprobar cómo de precisa es la verificación de la voz del usuario (comando de activación) en función de los modelos de IA generativa utilizados.**

Index Terms- Privacy, Generative AI, Voice cloning, Smart Personal Assistant

Tipo de contribución: *Investigación original*

I. Introducción

En los últimos años, gracias a la expansión de la Inteligencia Artificial, se han desarrollado y mejorado diferentes esquemas de interacción entre las personas y los dispositivos electrónicos. Uno de los avances más populares es la interacción por voz, en la que los usuarios son capaces de comunicar instrucciones a un dispositivo mediante 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 es la de los asistentes personales inteligentes o Smart Personal Assistant (SPA) [2]. 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].

En trabajos anteriores se analizó la seguridad de asistentes personales utilizando voces grabadas [5], [6]. En este trabajo se analiza el nivel de seguridad y privacidad de varios asistentes personales utilizando diferentes tipos de voces. El mayor reto ha consistido en analizar la posibilidad de emular la voz del propietario del asistente personal, lo que supone poder enviar el comando de activación para posteriormente realizar cualquier tipo de acción. Hoy en día hay una creciente actividad en todo lo relativo a Inteligencia Artificial generativa, y más concretamente en el clonado de voz [7], [8], [9]. Entonces se han analizado diversos sistemas de emulación de voz y se han ensayado los modelos con 3 asistentes personales. En concreto se ha trabajo con los siguientes asistentes personales y versiones: Apple HomePod Mini 16.6, Amazon Alexa Echo Show 5 y Google Home Mini 356012.

II. Análisis de comandos y funcionalidades soportadas

Para comenzar, se han elegido los comandos que verificasen diferentes niveles de seguridad en los SPAs. El más básico de todos es pedir información del tiempo ya que debería poder ser activado por todas las voces sin hacer ningún tipo de verificación de voz (por ejemplo "What's the temperature?"). A continuación, se ha elegido el comando "who am I?" para establecer si el dispositivo ha sido capaz de identificar al usuario, y sería señal de que puede haber brechas de seguridad en otros comandos. Para probar comandos involucrando dinero se ha intentado comprar algo de Amazon, aunque el único dispositivo realmente capaz de realizar esta función ha sido Amazon Alexa. Por último, se han intentado comandos relacionados con la privacidad, es decir, pedir la dirección de casa, el número de teléfono, información de contactos de la agenda y realizar llamadas. Se considera que el riesgo de contestar a las preguntas va incrementando en peligro según el orden en que se han planteado.

Para establecer la seguridad inicial se han realizado todas las pruebas sin haber registrado ninguna voz. Óptimamente, el SPA debería contestar solo a la primera pregunta, ya que sin haber registrado ninguna voz y sin hacer ningún reconocimiento no debería revelar información privada ni usar dinero de la cuenta. Como se puede ver en Fig. 1, Apple HomePod es el único que ha hecho lo esperado en todas las opciones ensayadas. Amazon Alexa permitió realizar una compra en Amazon sin ningún tipo de verificación, ni siquiera mandando un mensaje al móvil. Además, ha desvelado el domicilio de la persona. Esto no es importante en sí, ya que el dispositivo suele estar en casa de la víctima, pero sí es señal de

una brecha de seguridad. Por último, Google Home Mini reveló información privada no solo sobre el usuario, sino también de sus contactos.

Posteriormente, se registró una voz en los tres dispositivos y con esa misma voz se volvieron a realizar las mismas pruebas para ver realmente qué comandos están permitidos una vez se ha realizado el reconocimiento de voz. Es curioso notar que Apple HomePod Mini prioriza verificar a la persona mediante otros dispositivos Apple, ya sea el teléfono, el ordenador o el reloj. En estos otros dispositivos se habilita Siri y le sirve al HomePod para reconocer al usuario. Para realizar todas las pruebas con el Apple HomePod Mini se han apagado los otros dispositivos Apple cerciorándose de que se realiza la verificación por reconocimiento de voz.

Con esta última prueba se han acotado los comandos que están permitidos en cada dispositivo con acceso completo. Se puede ver en la Fig. 1 un resumen de las pruebas descritas anteriormente. El color de fondo verde indica un funcionamiento correcto o esperado, mientras que el fondo rojo indica un comportamiento inadecuado desde el punto de vista de vista de seguridad o un comportamiento no esperado. El fondo gris se ha utilizado para indicar que el dispositivo no soporta la funcionalidad o la prueba no tiene sentido en esas condiciones. Por ejemplo, sólo Amazon Alexa puede realizar compras en Amazon mientras que los otros dispositivos sólo pueden realizar búsquedas de productos en la web, pero no la compra directamente.

		Apple Homepod Mini 16.6			Amazon Alexa Echo show 5		Google Home Mini 356012	
		No training: any voice	Authorized voice	Authorized voice + iPhone	No training: any voice	Authorized voice	No training: any voice	Authorized voice
Unlock + Standard command	Play music	responde	responde	responde	responde	responde	responde	responde
who am i		-	responde	responde	-	responde	-	responde
Send messages		no	no	responde	no	no	no	no
Commands involving money	Buy something from amazon	no	no	no	responde	responde	no	no
Commands involving privacy	Where is my home	no	no	responde	no	no	responde	responde
	my phone number	no	no	responde	no	no	responde	no
	personal info from contacts	no	no	responde	no	no	responde	responde
	call someone	no	no	responde	no	no	no	no

Fig. 1 Comandos y funcionalidades soportadas

III. Análisis con distintas voces humanas

En la siguiente fase de pruebas se han utilizado distintas voces, pero todas de personas reales. Se ha probado primero con las voces de otras personas que no son las registradas en el dispositivo y luego con grabaciones de la voz con la que se había configurado el dispositivo. Por último, se ha buscado averiguar cuándo se realiza la verificación de voz en los dispositivos. Para ello se han realizado comandos con dos personas, donde una de ellas es la que tiene la voz registrada. Se han probado los distintos comandos con ambas voces intercambiando cuánto dice cada persona para ver hasta qué punto se verifica la voz y si hay comandos en los que verifique más palabras porque sean más privados.

El resultado esperable de estas pruebas sería que los dispositivos sólo respondan a comandos inofensivos, como sería el caso de no haber registrado ninguna voz. Sin embargo, estas pruebas han permitido averiguar que todos los dispositivos realizan la validación de la voz con el comando de activación ("Hey Siri" ó sólo "Siri", "Hey Google", "Alexa"), permitiendo que cualquier otra voz indique el comando a ejecutar. Para las pruebas de envío de mensajes con Apple HomePod es necesario que el iPhone esté conectado y en rango de alcance, ya que es el teléfono quien se encarga del envío del SMS. A continuación, se muestran los resultados de estas pruebas donde el color verde indica la voz del propietario del dispositivo, azul la voz de otro usuario, y en negro la respuesta obtenida.

Hey Siri, who am I? You are Clara
Siri, who am I? You are Clara

Hey Siri, who am I? I'm not sure who's speaking
Siri, who am I? I'm not sure who's speaking

Hey Siri, send a message to xxxxx saying Hi --> ... Sent
Siri, send a message to xxxxx saying Hi --> ... Sent

Hey Siri, send a message to xxxx --> Pide autenticación

Hey Google, who am I? You are Clara
Hey Google, who am I? I can only share personal information with

Hey Google, what is xxx's email? --> Desvela el email de xxx
Hey Google, what is xxx's email? --> No admite el comando

Alexa, who am I? You are Clara
Alexa, who am I? Don't know who you are, but this device is configured for Clara

En general, el comportamiento es el esperado, una vez que se ha descubierto que la validación tiene lugar analizando el comando de activación. Entonces, si el comando de activación lo dice el propietario del dispositivo se puede realizar a continuación cualquier otra acción independientemente del tipo de voz, mientras que si el comando de activación lo dice otra persona ya no es posible ejecutar comandos sensibles. En el caso de envíos de mensajes de texto, que es una funcionalidad propia del entorno Apple, el iPhone debe estar encendido y al alcance del Apple HomePod y cuando la voz de activación no es la correcta da la opción de validarse con el iPhone, ya sea mediante Face ID o mediante passcode. Es decir que el sistema es seguro, pero si hay un ambiente de ruido, o la persona está afónica, o existe cualquier otro problema con la validación de la voz, se puede realizar la validación en el teléfono que le pasa el visto bueno al Apple HomePod.

A continuación, se han realizado las pruebas con grabaciones de todos los comandos analizados de la persona que tiene la voz registrada. Todos los mensajes se han grabado en las mismas condiciones y se han reproducido al mismo nivel de volumen y misma distancia de los tres dispositivos. Los tres dispositivos han sido engañados con el comando “who am I?”, manteniendo el resto de su comportamiento igual que en la prueba anterior. Desvelar el nombre del propietario del dispositivo ante el comando “who am I?” es la prueba más sencilla de que el dispositivo ha validado la voz registrada, puesto que este comando dicho por una persona no registrada devuelve respuestas que rechazan el comando.
Sin embargo, en los primeros ensayos con voces grabadas uno de los dispositivos no respondía a la pregunta. Se ha detectado

que existe algún tipo de protección contra voces grabadas en los dispositivos Apple. Aunque se consigan grabar los comandos de activación de la persona registrada en ambientes libres de ruido, una reproducción normal no consigue activar Apple HomePod. Es conocido que la grabación digital y los algoritmos de compresión con pérdida de información no consiguen el nivel de calidad de sonido de las tecnologías de audio tradicionales [10]. Intuimos que este asistente pone su foco en contenidos de alta frecuencia que fácilmente se atenúan en el proceso de grabación, compresión o reproducción cuando no se cuenta con equipos especiales de alta calidad de sonido. Sin embargo, al reproducir los mensajes de voz de manera algo acelerada, como 1.2x o 1.5x, sí resultó posible activar el dispositivo y ejecutar comandos comprometidos en Apple HomePod. También se ha comprobado que ninguno de los dispositivos responde a los comandos grabados si se reproducen a 2x, ver Tabla 1.

Tabla 1 Velocidades de los audios a los que responden los SPAs

	Apple Homepod	Amazon Alexa	Google Home
1x	No	Responde	Responde
1.5x	Responde	Responde	Responde
2x	No	No	No

Es cierto que resulta complicado, en la práctica, conseguir grabar el comando de activación por voz de una persona sin su colaboración. En un ambiente público siempre hay ruido de fondo, y no es posible situar el micrófono cerca de la víctima para registrar adecuadamente su comando de activación sin que se dé cuenta. Pero si se consigue grabar el comando de activación de una persona, entonces su Asistente Personal sería vulnerable, pues bastaría reproducir el comando de activación y a continuación solicitar muchas de las acciones. El resumen de las acciones que se pueden realizar con voz no autorizada y con comandos grabados se puede ver en la Fig. 2. Como se ha comentado anteriormente, en el caso de Apple HomePod muchos de los comandos requieren que el iPhone se encuentre cerca del dispositivo. Entonces, aunque la grabación permita pasar la comprobación del comando de activación, en la práctica, una persona no autorizada no obtendría la funcionalidad.

		Apple HomePod Mini 16.6		Amazon Alexa Echo show 5		Google Home Mini 356012	
		Unauthorized voice	voice reccording	Unauthorized voice	voice reccording	Unauthorized voice	voice reccording
Unlock + Standard command	Play music	responde	responde	responde	responde	responde	responde
who am i		no	responde	no	responde	no	responde
Send messages		no	iPhone	no	no	no	no
Commands involving money	Buy something from amazon	no	no	responde	responde	no	no
Commands involving privacy	Where is my home	no	iPhone	no	no	no	responde
	my phone number	no	iPhone	no	no	no	no
	personal info from contacts	no	iPhone	no	no	no	responde
	call someone	no	iPhone	no	no	no	no

Fig. 2 Resultados de ataques con distintas voces humanas

IV. Análisis con voces sintéticas

Las últimas pruebas realizadas son con voces sintéticas ya que, si se descarta la posibilidad de conseguir grabar con calidad el comando de activación del propietario del asistente personal, entonces cobra mucho interés la opción de generar voces sintéticas para acceder al dispositivo. Ya se ha comprobado que todos los dispositivos realizan una validación de la voz del propietario antes de ejecutar comandos comprometidos que desvelen información personal o que realicen acciones arriesgadas. Sin embargo, al sobrepasar el proceso de validación, cualquier persona o cualquier tipo de voz sintética podría ejecutar comandos de riesgo.

Para la primera prueba de voces sintéticas, se han utilizado sistemas text-to-speech, ya que son muy fáciles de generar, además de ser gratuitos y rápidos. El comportamiento esperado sería el mismo que cuando se utiliza una voz no registrada, es decir, que se puedan ejecutar comandos inofensivos como “play music”, pero que no se puedan ejecutar comandos comprometidos como “who am I?” o “send message to xxx”.

Los resultados, que se muestran en la siguiente tabla, indican que Apple HomePod ignora todas las órdenes recibidas del sistema text-to-speech. Aparentemente detecta que se trata de una voz sintética y simplemente ignora los comandos, incluidos los inofensivos. El resto de los dispositivos responde de la manera esperada, aceptando comandos inofensivos y rechazando los comprometidos. Amazon Alexa ha revelado menos información que en las pruebas anteriores con la voz registrada ya que no ha respondido a la pregunta “who am I?” ni ha contestado a la pregunta sobre el domicilio. Google Home Mini también se ha protegido frente a comandos comprometidos, no ha respondido a la pregunta “who am I?” y tampoco ha respondido a preguntas sobre el domicilio o el número de teléfono. Mediante estas pruebas se ha visto que la voz de text-to-speech es diferente a la del propietario y no consigue engañar a los SPAs. Adicionalmente, se ve que esta voz es significativamente distinta a la de los humanos ya que los dispositivos llegan a ignorarla por completo. En la Fig. 3 se muestran las respuestas a distintos comandos. Se ha marcado en fondo gris la *no respuesta* al comando “play music” porque cualquier tipo de voz debería poder activar la música, pero no hacerlo tampoco supone un peligro. El hecho de reconocer que se trata de una voz artificial activa un mecanismo general de seguridad que hace que el dispositivo ignore cualquier comando, incluso los inofensivos.

		Apple HomePod Mini 16.6	Amazon Alexa Echo show 5	Google Home Mini 356012
		using text to speech	using text to speech	using text to speech
Unlock + Standard command	Play music	no	responde	no
who am i		no	no	no
Send messages		no	no	no
Commands involving money	Buy something from amazon	no	responde	no

Fig. 3 Resultados de ataque con voz sintética text-to-speech

A continuación, se han analizado algunos de los sistemas de generación de voz basados en Inteligencia Artificial generativa disponibles en el mercado. La gran ventaja de estos sistemas, desde el punto de vista conseguir acceso a un Asistente

Personal, es que se pueden encontrar algoritmos que entrenan con cualquier grabación de voz de la víctima. En principio es posible entrenar un sistema de voz basado en IA generativa utilizando la grabación sin ruido de una conferencia, como por ejemplo una TED Talk, o de una intervención en los medios de comunicación, o de una llamada telefónica. Una vez entregada al sistema, es posible generar cualquier mensaje de voz que incluirá los rasgos y matices de la voz de la víctima, lo que puede permitir suplantar al propietario del Asistente Personal. Otros algoritmos solo permiten ser entrenados con frases específicas lo que complica la generación de la voz. Aun así, se puede trocear una grabación limpia de las mencionadas anteriormente para construir las frases que solicita el programa.

Para generar la voz clonada se han probado varias aplicaciones web distintas buscando la que más se asemeja a la original. Se han intentado usar cuatro: PlayHT[11], Resemble.AI [12], Lovo [13] y Speechify [14]. Se han descartado otras como Murf Studio [15], IIElevenLabs [16] y Elai [17] por tener un coste elevado. Cada aplicación utiliza un método de entrenamiento que suele ser o una grabación limpia o leer frases predeterminadas. Speechify se ha entrenado con un audio sin ruido de un minuto y medio, ha sido un entrenamiento rápido, pero nada preciso. Además, requería pagar para usar la voz que ha generado, por lo que se ha descartado para hacer las pruebas. Con Lovo se ha intentado entrenar el modelo con el mismo audio limpio de minuto y medio, pero no lo aceptaba por lo que se recurrió a leer las frases específicas que proporciona la aplicación. El aprendizaje ha sido rápido, pero poco preciso. Comienza hablando lento y según continúa va cogiendo un ritmo normal. Esto no es un gran problema si el objetivo es generar un audio de varios minutos, pero en cambio lo que se busca en este caso es generar un par de palabras y eso lo hace tan lento que no engaña a nadie. Por esa razón se ha descartado a la hora de hacer las pruebas. PlayHT ha sido entrenado con el mismo audio de minuto y medio que Speechify, el entrenamiento ha sido rápido, pero tarda mucho en generar el texto deseado. Además, la generación ha sido muy precisa. Por último, Resemble.AI ha sido entrenado leyendo frases específicas ya que era la única opción. El entrenamiento ha sido muy lento pero rápido al generar las frases deseadas. Este requiere pago en función de los segundos que genera. Este ha sido el que más se ha parecido a la voz original. Todos estos resultados quedan resumidos en la Fig. 4.

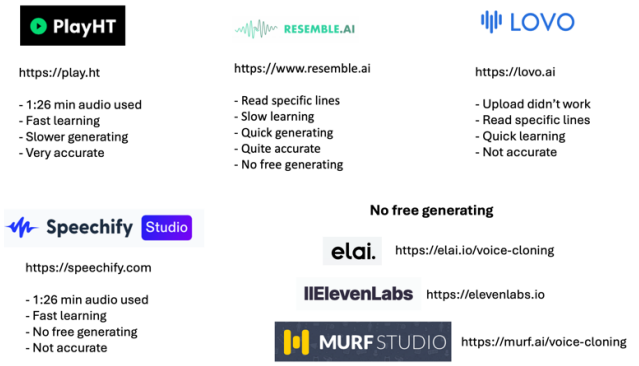

Fig. 4 Resumen de los modelos de IA analizados

Para verdaderamente probar si podían engañar a personas, se han hecho pruebas con personas que conocían a la voz original clonada. Se les ha proporcionado una grabación real y otras falsas sin decir cuantas había de las verdaderas o de las clonadas. PlayHT ha sido la única que ha engañado a personas reales. A la hora de probar estas voces con los SPAs, se ha utilizado únicamente el comando de activación, haciendo el resto del comando con otra voz. Se ha realizado de esta manera para demostrar que los dispositivos solo verifican el comando de activación y que con muy poca generación es posible acceder a toda la información como si fuese el propietario. En las pruebas se ha visto que Google Home Mini y Amazon Alexa han sido engañados con ambas voces. En cambio, el Apple HomePod no ha sido engañado con PlayHT, aunque con Resemble.AI ha respondido "you sound like xxxx", por lo que se puede deducir que, aunque haya sido engañado, la verificación es mucho más precisa que en los otros dispositivos ya que nota más detalles.

V. Resultados

Si no se registra ninguna voz en los asistentes, se ha comprobado que tienen unas funcionalidades muy limitadas ya que los únicos comandos permitidos son aquellos que no tienen riesgo, que son los denominados comandos inofensivos en este artículo, como pueden ser "play music" o solicitudes de información pública como la información meteorológica, o resultados de búsquedas en la web.

Con la configuración adecuada, que fundamentalmente consiste en registrar voces autorizadas en el asistente, se pueden realizar acciones más avanzadas o más delicadas. Sin embargo, se ha comprobado que todos los asistentes realizan la validación del usuario autorizado mediante el comando de activación. Eso significa que basta con suplantar dicho comando de activación para poder realizar cualquier acción.

En las pruebas con voces humanas reales, se ha comprobado que los SPAs ensayados sólo se activan con la voz del propietario. Sin embargo, en la mayoría de los casos se ha conseguido realizar la validación con grabaciones de la voz autorizada.

En el caso de voz sintética utilizando sistemas text-to-speech, no ha sido posible generar un comando de activación que el dispositivo valide correctamente. Esto es un resultado esperado ya que, como mucho, se puede considerar que la voz generada con estas herramientas sería equivalente a la voz de una persona no autorizada.

En Fig. 5, Fig. 6 y Fig. 7, se muestran las acciones que cada uno de los dispositivos (Apple HomePod, Amazon Alexa y Google Home Mini) ha ejecutado en función del tipo de voz utilizada.

		Apple HomePod Mini 16.6				
		No training: any voice	Authorized voice	Unauthorized voice	Voice reccording	Using text to speech
Unlock + Standard command	Play music	responde	responde	responde	no	no
who am i		-	responde	no	responde	no
Send messages		no	iPhone	no	iPhone	no
Commands involving money	Buy something from amazon	no	no	no	no	no
Commands involving privacy	Where is my home	no	iPhone	no	iPhone	
	my phone number	no	iPhone	no	iPhone	
	personal info from contacts	no	iPhone	no	iPhone	
	call someone	no	iPhone	no	iPhone	

Fig. 5 Resultados de todos los ataques a Apple HomePod

		Amazon Alexa Echo show 5				
		No training: any voice	Authorized voice	Unauthorized voice	Voice reccording	Using text to speech
Unlock + Standard command	Play music	responde	responde	responde	responde	responde
who am i		-	responde	no	responde	no
Send messages		no	no	no	no	no
Commands involving money	Buy something from amazon	responde	responde	responde	responde	responde
Commands involving privacy	Where is my home	responde	no	no	no	
	my phone number	no	no	no	no	
	personal info from contacts	no	no	no	no	
	call someone	no	no	no	no	

Fig. 6 Resultados de todos los ataques a Amazon Alexa

		Google Home Mini 356012				
		No training: any voice	Authorized voice	Unauthorized voice	Voice reccording	Using text to speech
Unlock + Standard command	Play music	responde	responde	responde	responde	no
who am i		-	responde	no	responde	no
Send messages		no	no	no	no	no
Commands involving money	Buy something from amazon	no	no	no	no	no
Commands involving privacy	Where is my home	responde	responde	no	responde	
	my phone number	responde	no	no	no	
	personal info from contacts	responde	responde	no	responde	
	call someone	no	no	no	no	

Fig. 7 Resultados de todos los ataques a Google Home

Por lo tanto, el acceso a comandos avanzados sólo es posible mediante la voz del propietario del dispositivo (real o grabada). Sin embargo, dado que se considera difícil conseguir grabar con buena calidad el comando de activación de una persona, el mayor esfuerzo de esta investigación se ha puesto en el análisis de herramientas de clonado de voz basadas en Inteligencia Artificial generativa disponibles en Internet. Utilizando los dos sistemas que mejores resultados han dado, se ha podido acceder a todos los dispositivos ensayados mediante la generación del comando de activación después de haber entrenado el generador de voz a partir de mensajes hablados que no incluían las palabras que forman parte de los comandos de activación. La Tabla 2 resume los resultados de estos sistemas de clonado de voz en los diferentes dispositivos. Apple HomePod no se ha activado con PlayHT, pero sí con Resemble.AI. Resulta curioso que en el caso de Resemble.AI, no responde "you are xxxx" sino "you sound like xxxx", como si no hubiera alcanzado un nivel de certeza suficiente en el proceso interno de validación de la voz.

Tabla 2 Resistencia a los ataques basados en IA generativa

	Play HT	Resemble AI
HomePod	NO	~ Responde
Alexa	Responde	Responde
Google Home	Responde	Responde

VI. Conclusiones

Con las pruebas realizadas en esta investigación se ha podido comprobar que los asistentes personales de las marcas más extendidas realizan un buen trabajo diferenciando las acciones que no entrañan riesgos frente a comandos que desvelan información personal o comandos que realizan acciones específicas. Cuando se solicitan acciones inofensivas como poner música o consultar la temperatura, los dispositivos admiten cualquier voz sin necesidad de validación. Sin embargo, cuando se solicitan datos personales o realizar acciones como enviar mensajes o realizar llamadas, todos los dispositivos requieren un proceso de validación de la voz del usuario. Se ha demostrado con las pruebas realizadas que dicha validación se basa en el reconocimiento de rasgos personales del comando de activación ("Hey Siri", "Hey Google", o "Alexa").

Las pruebas experimentales presentadas en este artículo demuestran que la utilización de voces sintéticas supone un riesgo de acceso a los Asistentes Personales. Con sistemas text-to-speech sólo se pueden realizar acciones inofensivas porque la voz que se obtiene es genérica y no incluye rasgos característicos del propietario. Con voces grabadas del usuario registrado sí ha sido posible ejecutar comandos delicados, sorteando las protecciones de Apple HomePod y Google Home Mini contra voces grabadas. El mayor éxito en el ataque a los Asistentes Personales se ha conseguido al aplicar IA generativa para imitar la voz del usuario registrado. Los modelos de IA han permitido simular los comandos de activación con rasgos característicos de la voz del propietario de los dispositivos, lo que nos ha permitido ejecutar comandos delicados. El riesgo que entraña este método de ataque es que el modelo de IA generativa se entrena con mensajes de voz del propietario del dispositivo que no incluyen ninguna de las palabras que forman parte de los comandos de activación.

Referencias

[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," *Proc ACM Interact Mob Wearable Ubiquitous Technol*, vol. 1, no. 4, pp. 1–23, Jan. 2018, doi: 10.1145/3161187.

[2] "Number of voice assistants in use worldwide 2019-2024 | Statista." Accessed: Mar. 15, 2024. [Online]. Available: https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/

[3] "Amazon Alexa – Learn what Alexa can do | Amazon.com." Accessed: Mar. 15, 2024. [Online]. Available: https://www.amazon.com/b?ie=UTF8&node=21576558011,

[4] "Smart speakers: use case frequency U.S. | Statista." Accessed: Mar. 15, 2024. [Online]. Available: https://www.statista.com/statistics/994696/united-states-smart-speaker-use-case-frequency/

[5] C. Valero *et al.*, "Evaluando la seguridad y privacidad de los asistentes personales inteligentes: ¡Ojo con el juguete!," in *VII Jornadas Nacionales de Investigación en Ciberseguridad - JNIC 2022*, Bilbao, 2022.

[6] C. Valero *et al.*, "Analysis of security and data control in smart personal assistants from the user's perspective," *Future Generation Computer Systems*, vol. 144, pp. 12–23, Jul. 2023, doi: 10.1016/J.FUTURE.2023.02.009.

[7] O. A. Shaaban, R. Yildirim, and A. A. Alguttar, "Audio Deepfake Approaches," *IEEE Access*, vol. 11, pp. 132652–132682, 2023, doi: 10.1109/ACCESS.2023.3333866.

[8] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. Kinnunen, "Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data," *Speaker and Language Recognition Workshop, ODYSSEY 2018*, pp. 240–247, Mar. 2018, doi: 10.21437/Odyssey.2018-34.

[9] H. T. Luong and J. Yamagishi, "NAUTILUS: A Versatile Voice Cloning System," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 28, pp. 2967–2981, 2020, doi: 10.1109/TASLP.2020.3034994.

[10] M. Buck, "Audio Quality: Performance Testing of Information Appliances." Audio Engineering Society, Mar. 01, 2001.

[11] "AI Voice Generator: Realistic Text to Speech and AI Voiceover | PlayHT." Accessed: Mar. 15, 2024. [Online]. Available: https://play.ht/

[12] "AI Voice Generator with Text to Speech and Speech to Speech." Accessed: Mar. 15, 2024. [Online]. Available: https://www.resemble.ai/

[13] "AI Voice Generator: Realistic Text to Speech & Voice Cloning." Accessed: Mar. 15, 2024. [Online]. Available: https://lovo.ai/

[14] "AI Voice Generator, Text To Speech, #1 Best AI Voice." Accessed: Mar. 15, 2024. [Online]. Available: https://speechify.com/

[15] "AI Voice Cloning Online: Clone Your Voice in Seconds." Accessed: Mar. 15, 2024. [Online]. Available: https://murf.ai/voice-cloning

[16] "AI Voice Generator & Text to Speech | ElevenLabs." Accessed: Mar. 15, 2024. [Online]. Available: https://elevenlabs.io/

[17] "Elai.io - AI Voice Cloning: Clone Your Voice Effortlessly." Accessed: Mar. 15, 2024. [Online]. Available: https://elai.io/voice-cloning