Journal of Information Security and Applications 101 (2026) 104554

Contents lists available at ScienceDirect

Journal of Information Security and Applications

journal homepage: www.elsevier.com/locate/jisa

Analysis of the security and privacy of smart personal assistants with real and synthetic voices

Clara Palacios-Castrillo [a,b,*], Rafael Palacios [b,c], Roberto Gesteira-Miñarro [b], Alejandro Chávez-Macías [b], Gregorio López [b]

[a] Carnegie Mellon University, Department of Electrical and Computer Engineering. 5032 Forbes Avenue, Pittsburgh, PA, 15289, USA
[b] Universidad Pontificia Comillas, Instituto de Investigación Tecnológica. Alberto Aguilera 23, Madrid, 28015, Spain
[c] Massachusetts Institute of Technology. Cybersecurity at MIT Sloan. 77 Massachusetts Avenue, Cambridge, MA, 02139, USA

ARTICLE INFO

Keywords:
Privacy
Generative AI
Voice cloning
Smart personal assistant
Cybersecurity
DeepFake voices

ABSTRACT

Smart Personal Assistants (SPA) can be trained with the owner's voice, and its voice features act as a biometric access password. The aim of this work was to analyze what information different personal assistants reveal without verifying the owner's voice, and what real risks exist in impersonating the owner's voice. To do this, a test protocol was defined, including commands for demanding generic information, personal information, and more sensitive requests such as making calls or purchases. To deceive the personal assistants, tests were carried out with various synthetic voices, including generative AI systems to create voice models based on the user registered in the assistants, hence allowing commands to be synthetically generated with the person's voice features. This study worked with Apple HomePod, Amazon Alexa, and Google Home assistants, which are the main devices on the market. It was possible to verify what type of information each system communicates without performing user validation and how accurate was the voice verification algorithm (activation command) depending on the synthetic voices used. We proposed a Synthetic Speech Detection system as a secondary security layer to identify whether a voice mimicking a target individual was synthetically generated. To evaluate this, a preliminary study on the fidelity of modern synthetic voices was conducted through subjective listening tests. The results indicate that human participants attained only a marginal performance above the 50% stochastic baseline, confirming the high perceptual transparency of current models and the inherent difficulty of the detection task.

1. Introduction

In recent years, thanks to the expansion of artificial intelligence, various interaction schemes between humans and electronic devices have been developed and improved. One of the most popular advances is voice interaction, which allows users to give instructions to a device using voice commands, making interaction faster and more natural [1].

One of the technologies that has most contributed to improving voice interaction systems is that of intelligent personal assistants or Smart Personal Assistants (SPA) [2]. SPAs allow a user to perform many tasks using their voice as a means of transmitting information. For example, SPAs provide the ability to check information such as the weather or traffic conditions, listen to music, make voice and video calls, shop online, or control other devices such as smart lights and thermostats [3,4]. In contrast with other devices such as tablets or smart phones, SPAs are used exclusively by means of voice command, without any need to grab the device or type any commands. That is one of the characteristics that make them so accessible, intuitive and easy to use. However, the ability to use the device in the distance imposes access control limitations because they don't have a keyboard to type a password, or a fingerprint reader, or camera for face identification; any user validation should be done by means of voice identification.

If these devices were limited to playing music, like a Bluetooth speaker, then access control would not be necessary. Nevertheless, these devices are internet connected, and local network connected, providing a variety of functionalities, some of which could be considered delicate because of privacy or security concerns. The focus of this paper is to analyze the level of voice authentication necessary to access different functionality, organizing the tasks in increasing order of privacy and security risk.

* Corresponding author.
E-mail addresses: cpalacio@andrew.cmu.edu, clara.palacios.spain@gmail.com (C. Palacios-Castrillo).

https://doi.org/10.1016/j.jisa.2026.104554

Available online 29 June 2026

The methodology used in this work involves three major steps. First, we tested basic functionality of the devices; this is described in Section II. Second, we analyzed the behavior using human voice, the voice of the owner of the device and other people (Section III). And third, we used synthetic voices generated with different methods (Section IV). In view of the results, and the threat of AI-based voice cloning, we discuss about ways to distinguish synthetic and human voices.

Several voice cloning systems have been analyzed in this work, and some AI models have been tested with the three top of the market personal assistants. Specifically, work was done with the following personal assistants and versions: Apple HomePod Mini 16.6, Amazon Alexa Echo Show 5 (3rd Gen), and Google Home Mini 356,012.

2. Background and motivation

Previous work has analyzed the security of personal assistants using recorded voices [5] and security and privacy of wearable devices [6,7]. In this work, the level of security and privacy of several personal assistants is analyzed using different types of voices. The main challenge was to analyze the possibility of emulating the voice of the owner of the personal assistant, which would mean being able to send the activation command and then perform any action.

Today, there is a growing activity in the field of generative artificial intelligence. One of the fields of generative AI is the creation of multimedia content, denoted as DeepFake, that applies to Video, Images and Audio. An extensive review of methods for creating (and also detecting) audio deepfakes can be found in [8]. The quality of the audio created with the newest Generative Adversarial Networks (GAN) is so realistic, that now a days it is very difficult to distinguish between real audios recorded by human and audios synthetically generated. In this paper we have also included an analysis about the limitations for human to distinguish fake audios generated with the latest technology and real human audios.

The advances in Generative AI, allows even to apply deepfake voice to perpetrate social engineering attacks, in a similar way as phishing attacks trick users by email. The new AI-based voice generation techniques produce such a "natural" voice and so quickly that they can be used to manipulate humans through voice phone calls, in a technique called voice phishing or vishing [9]. In fact, deepfake voices can be used in many different scam schemes, including the scam the CEO of a company [10].

More specific to this research is the possibility to synthetically generate voice with the characteristics of a particular individual, allowing the attacker to impersonate that individual. This specific type of AI generation of voice to impersonate someone is called speech cloning or voice cloning [11–13]. Previous research has also analyzed the use of audio deepfakes attacks to personal devices [14] with sophisticated deep learning frameworks such as Real Time Voice Cloning RTVC and Speaker Verification to Multi-speaker Text-to-Speech Synthesis (SV2TTS) based on Tacotron [15]. VoiceBox [16], developed by Meta, is a high-quality AI model for text-to-speech synthesis that even includes speech editing, noise removal and multilingual capabilities. Although the model could easily be used for voice cloning, Meta has not publicly released the code nor made the model accessible for users. Particularly because of the concerns about malicious uses of the model, highlighting the risks of voice impersonation, deepfake generation, and identity fraud. So VoiceBox was not used in this research due to the restrictions imposed by the developer.

However, more threatening is the fact that nowadays many of these tools no longer require advanced knowledge or large computational resources, since they are available on the internet for everyone to use. Maybe not the state of the art in synthetic voice generation or voice cloning, but in this work, we wanted to evaluate if those highly available tools really suppose a threat to the most popular SPAs. Jemine et al. [17] developed SV2TTS Toolbox able to clone voice based on Tacotron2, and made the code available on GitHub. Although not as accessible and easy to use as web-based services, it was considered sufficiently available to the public and was also included in the study. The most recent free online tool is called Voice Cloning Studio; it was developed by Hasan Basbunar, and is available at huggingface [18].

3. Analysis methodology

3.1. Testing environment and procedures

To replicate real-world conditions, we conducted the experiments in a home environment. The living room, measuring 6.3 by 4.1 m with a 2.6 m ceiling, features standard decorations like fabric curtains and a sofa that absorb reverberation. While it's not an anechoic chamber, it provides a realistic home environment for more authentic measurements but maintaining silent conditions to minimize noise interference during the experiments. The base noise level was measured using a digital sound level meter, Smart Sensor AS804. The sound level measured in the room had an average level of 31.4 dBA, with 97% of the samples between 30 and 33 dBA. This is a very low noise environment, especially compared to a sound level of 60 to 70 dBA while a person is speaking.

Audios for making tests directly on SPA or for training the models were recorded either at 44.1 kHz or 48 kHz. Nevertheless, the models may resample internally to the desired sample frequency; for example, SV2TTS is known to resample to 16 kHz using the librosa library. The audios generated by the models had a sampling rate of 44.1, 48 or 16 kHz, depending on the model, and they come in lossless format like frac or lossy formats like MP3 or M4A.

Speaking or playing command to the Smart Personal Assistants, was done within a 1 m distance. Many tests were carried out, especially while playing commands. Different types of speakers, volume level, and audio player applications. We found that SPAs were very reliable and consistent. With human voice command, only one out of ten times the device didn't receive the message. In the case of synthetically generated audios, a few additional adjustments were required. When the device couldn't "hear" or "understand" the voice, it simply ignored the message or, in some cases, responded with messages like "I don't understand what you said" or "sorry, try asking again." However, when the message reaches the device, it either executes the command demanded or refuses to do so. Typical replies to commands by an unauthorized person include "I can only do that once I verify your voice" or "I'm not sure who's speaking." As an example, the answer to "who am I" asked by an unauthorized voice would be "You are a person" instead of "You are [NAME]."

The crucial point is that the behavior is highly deterministic. If the message is not clearly received by the device, it is mostly ignored, and the experiment must be repeated. However, when the device receives the message, it either executes the command if authorized, answering with a message that confirms the execution, or refuses to execute if not authorized, answering with a rejection message. In order to demonstrate the deterministic behavior of the devices we conducted more that 100 additional tests using both real voices and cloned voices. Across all tests, we did not find any voice–device combination that behaved inconsistently (i.e., accepting a voice in some trials and rejecting it in others). While noisy environments might potentially introduce variability, all our experiments were conducted under controlled conditions, and results were fully consistent.

Running additional repetitive tests long enough we were able to find a few discrepancies. We used the main authorized human voice, and two cloned voices using XTTS model (one male, one female). Running 50 tests with each voice, we found that all human trials were correctly identified (100% accuracy), only 2 of female tests were wrongly identify (98% accuracy), and only 1 of the male tests was misclassified (99% accuracy). These results indicate that the authentication mechanism maintains a high degree of empirical consistency, as the vast majority of synthetic attempts were continuously rejected while legitimate access

remained perfectly validated. Given the high degree of empirical consistency where results yielded near-zero variance across all trials, a binary 'Red/Green' visualization was adopted to represent the outcomes.

3.2. Command and functionality analysis

To start, commands were chosen to verify different levels of security and privacy in SPAs. The most basic of all is asking for the weather, as it should be activated by all voices without any kind of voice verification because it doesn't impose any threat (a similar command will be "What's the temperature?"). Then, the command "who am I?" was chosen to establish whether the device could identify the user, which would be a sign that there could be security breaches in other commands. To test commands involving money, an attempt was made to purchase something from Amazon, although later we found that the only device capable of doing this was Amazon Alexa. Finally, commands related to privacy were attempted, such as asking for the home address, phone number, contact information, and making phone calls. It is considered that the risk of answering the questions increases in danger according to the order in which they were posed. These initial results are shown in this section because they are the basis for defining the subsequent tests.

To establish the initial security with the default configuration (out of the box settings), all tests have been performed without registering any voice as owner of the SPA. Optimally, the SPA should answer only the first question, since without having recorded any voice and without doing any recognition it should not reveal private information or use money from the account.

As can be seen in Fig. 1, Apple HomePod is the only SPA that has done as expected in all tested options. Amazon Alexa allowed to make a purchase on Amazon without any verification, not even sending a message to the cell phone. In addition, it reveals the person's home address. This is not important in itself, since the device is usually at the victim's home, but could be revealing a second-home address or the billing address instead, anyway it is a sign of a privacy breach. Finally, Google Home Mini revealed private information not only about the user, but also about his contacts. In graphs of results, we have used several icons to simplify the contents and make them more compact. Icons were created by various independent designers or companies: Raphaël Buquet, Rainbow Designs, Nice Design, Berkah, Ahmad Roaayala, Alvida, Paola Moreira, and Icon Depot, all available at the Noun Project [19]. Icons similar to the logos of Apple, Amazon and Google were used to identify the devices, and several icons were used to indicate the kind of action or command. The action icons on Fig. 1 (and subsequent graphs) correspond, from left to right, to the following actions: Play music, Who am I, send message, Online shopping, Where is my home, My phone number, Personal information of contacts, and Call someone.

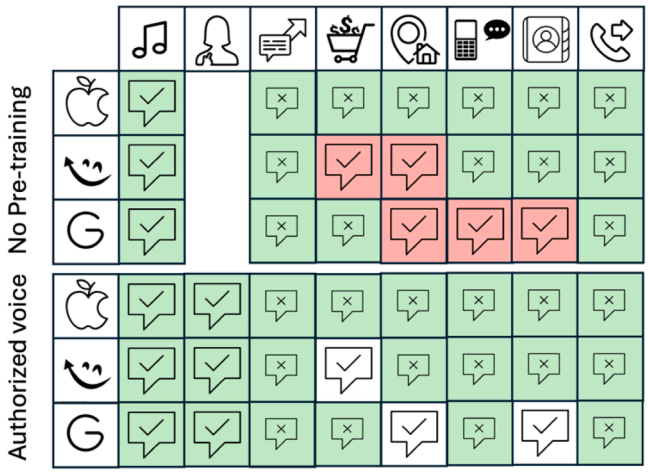

Fig. 1. Commands and supported functionalities.

The reaction of the device to each command is shown with an icon that indicates either an answer / positive reaction or not answer / rejection message. The green background color indicates correct or expected operation, from the security and privacy point of view, while the red background indicates inadequate behavior from this perspective. The white background has been used to indicate that the reaction is up to discussion, for example shopping without additional verification, or additional confirmation in another device. Only a few test were left without any response icon, because owner verification does not make sense if the device has not been trained yet with the authorized voice.

Afterwards, a voice was registered in the three devices, and the same tests were performed again with that same voice to see which commands become available once voice recognition has been performed. It is interesting to note that Apple HomePod Mini prioritizes verifying the person through other Apple devices, whether it be the phone, computer, or watch. Siri is enabled on these other devices and serves the HomePod to recognize the user. Therefore, to perform all tests with the Apple HomePod Mini, other Apple devices were turned off, ensuring that only voice recognition was performed without a second validation, like the rest of the devices.

With this last test, the commands allowed on each device with full access have been limited. A summary of the tests described above can be seen in Fig. 1. None of the devices can send text messages or make phone calls by themselves, since they need to make use of a nearby phone to operate on the mobile network. For Apple HomePod message-sending tests, the iPhone needs to be connected and within range as the phone is responsible for second validation and sending the SMS. Since iMessages can be sent through the Internet, in contrast with text messages that rely on SMS protocols and SIM cards, theoretically it is possible to send iMessages directly from the HomePod, however the extra layer of security prevented this if there isn't a nearby device for second validations. This requirement of second validation does not apply to iPads of Mac computers, which can send iMessages anytime since they all implement their own authentication mechanisms by keyboard, fingerprint, or face identification.

Regarding shopping options, only Amazon Alexa can make purchases on Amazon, while the other devices can only search for products on the web but not make the purchase directly because the use of native app is needed. It was quite surprising that Alexa allows to make purchases even without pre-training the voice. It's also a bit questionable it the authorized voice should be allowed to make purchases without additions verification, at least to avoid unintentional purchases. However, the products will always be sent to the registered owner's address, and the products can be returned, so it is more of a marketing decision to allow purchases so easily. It is also questionable if Google Home should be revealing personal information about the contacts. Considering that only the activation messages is verified, as it is explained below, once the device is unlocked anyone could extract private information.

3.3. Analysis with different human voices

In the next phase of tests, different voices were used, but all were real people. First, the voices of other people who were not registered on the device were tested, and then recordings of the voice with which the device had been configured were used. Finally, it was sought to find out when voice verification is performed on the devices. For this purpose, commands were performed with two people, where one of them has the registered voice. Various commands were tested with both voices, switching who says what to see to what extent the voice is verified and if there are commands where more words are verified because they are more private.

The expected result of these tests would be that the devices only respond to harmless commands as would be the case if no voice had been registered. However, these tests have revealed that all devices perform the voice validation with the activation command ("Hey Siri" or just "Siri", "Hey Google", "Alexa"), allowing any other voice to indicate the

command to be executed. On Fig. 2 are the results of these tests, where green indicates the device owner's voice, blue indicates another user's voice, and black indicates the response obtained.

In general, the behavior is as expected once it was discovered that validation takes place by analyzing the activation command. So, if the activation command is said by the device owner, any other action can be performed regardless of the type of voice, whereas if the activation command is said by someone else, sensitive commands cannot be executed. As explained before, in the case of sending text messages, which is a functionality specific to the Apple environment, the iPhone must be turned on and within reach of the Apple HomePod, and when the activation voice is incorrect, it gives the option to authenticate with the iPhone, either through Face ID or passcode. That is to say that the system is secure, but if there is a noisy environment, or the person is hoarse, or there is any other problem with the validation of the voice, authentication can be performed on the phone that transfers a passcode to the Apple HomePod.

Subsequently, tests were carried out with voice recordings of all the commands analyzed of the person who has the registered voice. All messages were recorded under the same conditions and played at the same volume level and distance from the three devices. However, in the initial tests with recorded voices, one of the devices did not respond to the question. It was speculated that there is some protection against recorded voices in Apple devices. Even if an attacker manages to record the activation commands of the registered person in noise-free environments, normal playback fails to activate Apple HomePod. It is known that digital recording and lossy compression algorithms do not achieve the same sound quality as traditional audio technologies [20]. We suspect that this assistant focuses on high-frequency content, which is easily attenuated in the recording, compression, or playback process when not using special high-quality sound equipment. However, we found that when the voice recordings are played slightly sped up, like 1.2x or 1.5x, it was possible to activate the device and execute compromised commands on Apple HomePod. It was also found that none of the devices respond to recorded commands when played at 2x speed.

3.4. Analysis with synthetic voices

The last tests carried out were with synthetic voices since if the possibility of obtaining a high-quality recording of the personal assistant owner's activation command is ruled out in practice, then the option of generating synthetic voices to access the device becomes very interesting. It has already been shown that all devices perform voice validation before executing compromised commands that reveal personal information or perform risky actions. However, once the validation process is bypassed, any person or any type of synthetic voice could execute risky commands.

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 --> Asks for authentication

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? --> Reveals xxx's email
Hey Google, what is xxx's email? --> Command not accepted

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

Fig. 2. Example of conversation using different human voices. Green=Authorized voice, Blue=another voice, Black=device answer.

For the first synthetic voice test, text-to-speech systems were used, as they are very easy to generate, as well as being free, hence highly available to any potential attacker. The expected behavior would be the same as when using an unregistered voice, meaning that harmless commands such as "play music" could be executed, but compromising commands such as "who am I?" or buy something, cannot be executed.

Next, we tested the systems using synthetic voices generated with Artificial Intelligence. As mentioned in the introduction, there is a lot of activity in the field of Generative Artificial Networks and voice cloning systems. We focused our research of the robustness of SPAs against generative AI by using AI-based voice generation systems available on the market, since these systems are already in the hands of any attacker. A great advantage of these systems, from the point of view of gaining access to a Personal Assistant, is that some algorithms can be trained with any recording of the victim's voice. In principle, it is possible to train a generative AI voice system using a noise-free recording of a conference such as a TED Talk or a media appearance or a phone call. Once delivered to the system, any voice message can be generated, including the characteristics and nuances of the victim's voice, which may allow impersonation of the personal assistant's owner. Other algorithms only allow training with specific phrases, complicating voice generation. Even so, a clean recording like the ones mentioned earlier can be split to construct the phrases requested by the program.

To generate the cloned voice, various web applications were tested, looking for the one that most resembles the original voice. Four were attempted: PlayHT [21], Resemble.AI [22], Lovo [23], and Speechify [24]. Others like Murf Studio [25], IIElevenLabs [26], and Elai [27] were discarded due to high cost. In addition to web-based systems, SV2TTS Toolbox [17] was also evaluated because it was considered easily accessible because the code is available and only requires local installation.

Each voice-cloning application uses a training method that is usually a clean recording or reading predetermined phrases. Speechify was trained with a noise-free audio of one and a half minutes, the training was quick, but the resulting voice was not accurate at all. Additionally, it required payment to use the generated voice, so it was discarded for testing. With Lovo we tried to train the model with the same minute and a half clean audio, but it did not accept it, so we resorted to reading the specific phrases provided by the application. The training was quick but the results not precise. The generated voice starts talking very slowly and picks up a normal pace as it continues. This is not a big problem if the goal is to generate an audio of several minutes, but instead what is sought in this case is to generate a couple of words and this application makes the speed of the voice so slow that it did not sound realistic and for this reason, it was also discarded for further testing. PlayHT was trained with the same one-and-a-half-minute audio as Speechify. The training was quick, and although it took a long time to generate the desired text, the synthetic voice was very precise and realistic. Resemble.AI was trained by reading specific phrases because the model is not prepared for free speech training, and those fixed messages were the only viable options for training. The training was very slow, but the model was quick in generating the desired phrases. Resemble.AI requires a small payment based on the seconds it generates. All these voice cloning systems are summarized in Fig. 3.

Tacotron is and end-to-end neural network architecture for Text-To-Speech (TTS) developed by Google and introduced in 2017 [15] that was later improved in 2018 in collaboration with Berkeley [28]. Tacotron takes characters as input and produces synthesized speech. It uses a module called CBHG that includes a Convolution Bank, followed by Highway networks and a bidirectional Gated recurrent unit, for extracting representations from sequences. Then it uses an encoder-decoder architecture with RNN attention layers to finally

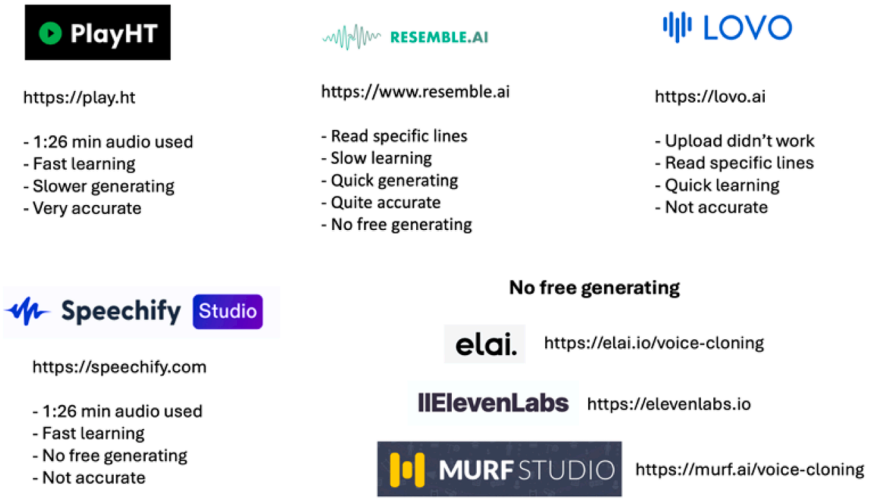

Fig. 3. Summary of the analyzed web-based AI models.

produce spectrograms that are fed to the Griffin-Lim reconstruction algorithm to synthesize the voice corresponding to the input characters [15]. Tacotron2 improved the process by mapping character embeddings to mel-scale spectrograms, followed by a WaveNet model acting as a vocoder to converts those spectrograms into audible audio waveforms.

Fine-tuning a speech model like Tacotron2 to target a specific speaker is a challenging task due to the substantial amount of voice data required to capture all linguistic tokens (the diverse speech sounds) necessary for the model to generate any possible voice message. Even with lightweight systems to minimize the volume of data samples [29] good speech quality and similarity is attained with 100 specific sentences of data. However, SV2TTS Toolbox [17] makes voice cloning based on Tacotron using just a few seconds of voice. It uses a speaker encoder to extract embeddings specific to the speaker that are feed into a modified version of Tacotron2 to generate a mel-spectrogram corresponding to the desired sentence and finally uses a vocoder to obtain the audible sentence. It only requires a few seconds of the target voice to generate an audible sentence with the characteristics of the target person's voice. Code available at GitHub [30]. Although SV2TTS requires just 5 s of reference speech [17] we used a recording of 13 s to provide more diverse sounds. For the final synthetic voice generation, the toolbox can use Griffin-Lim vocoder or a more advanced WaveRNN vocoder. Griffin-Lim is very fast but produces a much lower quality voice that sounds awkward to the human ear, while WaveRNN produces realistic voices but requires longer computing time, which is not a problem in this case. Both algorithms were tested with all three devices.

The Voice Cloning Studio was announced in March 2025 and updated in November 2025, with is the version available online at the time of this writing. This new tool is based on the voice cloning technology Coqui XTTS v2 [31]. It is available on the web as "Voice Cloning & Text-to-Speech App with XTTS v2" [18] and provides a user interface very easy to use, being able to generate voices in 17 difference languages, and any text, all while maintaining the characteristics of your reference voice. The website claims that a 3-seconds audio sample is enough to train the model, but we provided a 14 s audio sample that also allowed for a very fast training.

When testing these AI generated voices with the SPAs, only the activation command was used, with the rest of the commands being done with another voice. This was done to demonstrate that the devices only verify the activation command and that with very little voice generation, it is possible to access all the information as if it were the owner. In the results section we will find that Amazon Alexa was fooled with all AI generated voices. However, the Apple HomePod uses a more precise algorithm to authenticate voices; it was not fooled with PlayHT or SV2TTS or Coqui XTTS, but we got access with synthetic voices generated with Resemble.AI. Google Home Mini rejected SV2TTS messages, but gave access to PlayHT, Resemble.AI and Coqui XTTS.

4. Results

Results are organized according to the type of voices used for the different analyses. In all cases, the three commercial devices were used: Apple HomePod, Amazon Alexa, and Google Home.

Graphical representations are used to show the results with more clarity. As it was described before, the icon with check means that the device replied to request, by either executing the action requested or by answering with the response. The background color is shown green when the device performed the expected action, while red in case of a privacy or security threat. In a few cases the background is white because the action is open to interpretation and cannot be considered a clear green or red.

4.1. Results for basic functionality

If no voice is registered in the assistants, it has been shown that they have very limited functionality as the only allowed commands are those that are not risky, which are called harmless commands in this article, such as "play music" or requests for public information like weather information or web search results.

With the proper configuration, which fundamentally involves registering authorized voices in the assistant, more advanced or sensitive actions can be performed. However, it has been shown that all assistants validate the authorized user through the activation command. This means that it is enough to impersonate that activation command to be able to perform any action on the device.

4.2. Results for human voices

In the tests with real human voices, it was shown that the tested SPAs only activate with the owner's voice. However, in most cases, it was possible to perform validation with recordings of the authorized voice. All three devices were fooled by the "who am I?" command, maintaining the rest of their behavior the same as in the previous test. Revealing the name of the owner of the device at the command "who am I?" is the simplest proof that the device has validated the recorded voice, since this command requested by an unregistered person returns responses that reject the command.

One may expect some type of protection against recorded activation commands but even the HomePod was easy to fool by playing the recorded sound at 1.5x speed (as explained in Section 3.3). The fact that

HomePod failed to play music with the authorized voice recording is unexpected because even the unauthorized voice can do it. This is more an incidental result because of the speculated protection about recorded voices and the attempts to circumvent such protection by playing at different speed, bit in any case it does not impose a threat.

Recording someone's activation command and using it to gain access to the device relies on the category of Replay Attacks. Nonetheless, it is difficult in practice to record a person's voice activation command without their cooperation. In a public environment, there is always background noise, and it is not possible to place the microphone close to the victim to properly record their activation command without them noticing. But if it were possible to record a person's activation command, then their Personal Assistant would be vulnerable as it would suffice to play the activation command and then request many of the actions. The summary of actions that can be performed with unauthorized voices and recorded commands can be seen in Fig. 4. As mentioned earlier, in the case of Apple HomePod, many of the commands require the iPhone to be close to the device. Therefore, even if the recording allows passing the activation command verification, an unauthorized person will not obtain the functionality in practice.

4.3. Results for synthetic voices

In the case of synthetic voice using text-to-speech systems, it was not possible to generate an activation command that the device would validate correctly. This is an expected result as the most that can be considered is that the voice generated with these tools would be equivalent to the voice of an unauthorized person.

The results shown in Fig. 5 indicate that Apple HomePod ignores all orders received from the text-to-speech system. It apparently detects that it is a synthetic voice and simply ignores the commands, including the harmless ones. The other devices responded as expected, accepting harmless commands and rejecting the compromised ones. Amazon Alexa revealed less information than in the previous tests with the registered voice, as it did not answer the question "who am I?" nor did it answer the question about the home address. Google Home Mini also protected itself against compromised commands, did not answer the "who am I?" question, and did not answer questions about the home address or phone number. These tests have shown that the text-to-speech voice is different from the owner's voice and cannot deceive the SPAs. Additionally, it shows that this voice is significantly different from that of humans, as the devices even ignore it completely. Recognizing that it is an artificial voice activates a general security mechanism that makes the device ignore any command, even the harmless ones. The action to play music was rejected in some cases, which was expected to be treated as any unauthorized voice but is not considered a threat and is therefore left white in the graph.

Fig. 4. Results of the analysis with human voices.

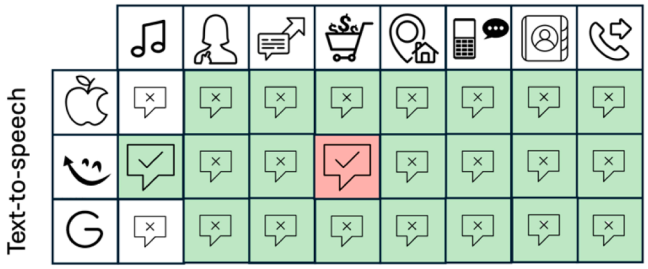

Fig. 5. Results of text-to-speech generated voices.

As it was expected, from the perspective of authentication, access to advanced commands is only possible through the device owner's voice. The voice of the owner of the SPA is used as a biometric identifier that replaces the most standard access control based on a secret password, since typing a password will not be very useful a human-to-machine interface centered on voice commands.

In most cases, a recording of the activation command can be used to unlock the device. However, since it is considered difficult to obtain a high-quality recording of a person's activation command, the greatest effort of this research has been focused on analyzing voice cloning tools based on generative artificial intelligence that are available on the internet. Using the system that gave the best results, it was possible to access all the tested devices by generating the activation command. It is important to note that AI models were trained with spoken messages that did not include the words that form the activation commands.

Fig. 6 summarizes the results of how these voice cloning systems interact with the different devices. Apple HomePod was not activated with PlayHT but it was activated with Resemble.AI. Interestingly, in the case of Resemble.AI it does not respond "you are xxxx" but "you sound like xxxx" when asked "who am I?", as if it had not reached a sufficient level of certainty in the internal voice validation process. The other SPAs, Alexa and Google Home, both activated and provided answers with the audios generated by PlayHT and Resemble.AI, as if the human owners of the devices were giving the commands. The audios generated by SV2TTS were rejected by HomePod and Google Home even in the high-quality version that uses WaveRNN. Only Amazon Alexa was fooled by SV2TTS and to the question "Who am I", it responded "You are [NAME], and this is [NAME]'s account". Surprisingly, when an unauthorized user gets physical access to an Alexa device and says "Alexa, who am I", the system answers "I'm not sure who's speaking, but you're in [NAME]'s account. To teach me to recognize your voice. Just say, learn my voice", and in fact it allows the unauthorized user to validate his/her voice and launch commands after the process.

As a summary, we present the results for the three SPAs analyzed with all types of voices. The following figures: Fig. 7, Fig. 8, and Fig. 9, show the actions that each of the devices (Apple HomePod, Amazon Alexa, and Google Home Mini) has executed depending on the type of voice used. It is really threatening, that using AI voice cloning the attacker performs a full validation and the behavior is the same as in the case of the Authorized voice.

4.4. Analysis of features of AI generated voices

In this section we present a comparative analysis of the legitimate

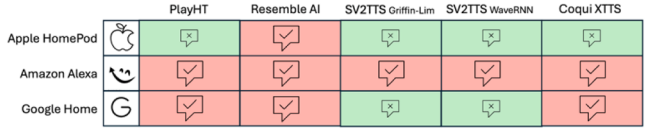

Fig. 6. Response against AI generated voice. Each cell represents the result of a minimum of 10 trials; see Section 3.1 for consistency validation.

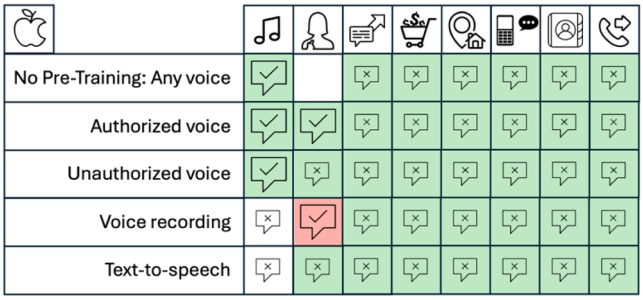

Fig. 7. Results of Apple HomePod.

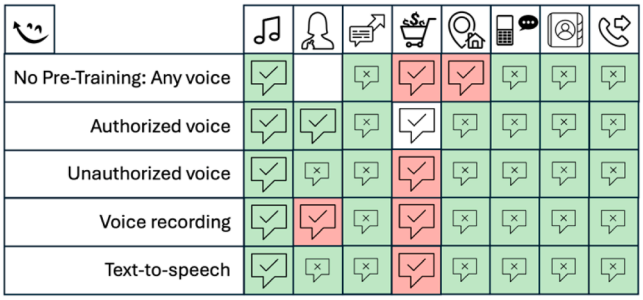

Fig. 8. Results for Amazon Alexa.

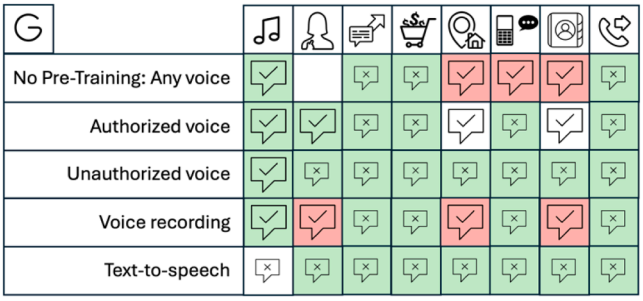

Fig. 9. Results for Google Home.

owner's original human voice and synthetic voices generated using two of the voice-cloning tools. We selected the three most representative audio samples of the phrase "Hey Siri, who am I": the original human recording, the voice generated using the ResembleAI tool and the voice produced by the latest Coqui XTTS v2 model. Although both AI-generated voices sound very realistic to the human ear, we previously demonstrated that only the sample generated with ResembleAI was able to authenticate successfully on the HomePod.

We computed key acoustic parameters (pitch, jitter and shimmer) for the three audio signals, as summarized in Table I. Additionally, we extracted Mel-frequency cepstral coefficients (MFCC) for spectral comparison, shown in Fig. 10.

Based on these results, it is difficult to determine which specific properties are most indicative of high-quality synthetic speech, or which properties are the key for a successful audio authentication. All three 1-second audio samples exhibit very similar MFCC patterns in Fig. 10, using a time resolution of 0.1 s. Regarding the acoustic parameters in Table I, ResembleAI sample is closer to Human voice in shimmer, but diverges more significantly in pitch and jitter. Conversely, the Coqui sample more closely matches the human recording in pitch but not in jitter or shimmer. Notably, both AI-generated audios are very similar to each other in jitter and shimmer, which aligns with prior studies related to the differences in prosody (represented by jitter and shimmer parameters) between genuine and fake speech [32].

Table I
Acoustic parameters of the 3 most relevant audio signals.

	Pitch (Hz)			
	Mean F0	Std F0	Min F0	Max F0
Human	294.0	23.5	237.9	337.9
ResembleAI	269.3	56.4	103.3	402.5
Coqui/XTTS-v2	255.1	23.1	222.0	302.3
	Jitter			
	Local %	RAP %	PPQ5%	
Human	0.020	0.010	0.011	
ResembleAI	0.014	0.006	0.007	
Coqui/XTTS-v2	0.016	0.008	0.009	
	Shimmer			
	Local %	APQ3%	APQ5%	
Human	0.061	0.022	0.030	
ResembleAI	0.056	0.017	0.027	
Coqui/XTTS-v2	0.071	0.028	0.038	

Another interesting observation is the proximity between the Coqui and Human samples in Pitch parameters. Recent work [33] has shown that pitch-based representations can support highly accurate deepfake detection. We can appreciate that the Coqui audio sounds very human-like. Nevertheless, smart personal assistants should prioritize identifying the unique voice characteristics of the legitimate owner rather than relying on human-like acoustic properties. In the next section, we argue that integrating deepfake detection mechanisms could provide an additional layer of security within the voice-based authentication process.

5. Discussion about synthetic voices

Smart Personal Assistants should evolve to include automatic detection of voices generated synthetically to avoid the types of attacks analyzed in this work. As it was shown before, SPAs currently do a good job analyzing voices generated with Text-to-Speech programs since they don't contain the unique characteristics of voice of the authorized owner of the device. Even in the case of voice recording, we found that Apple HomePod incorporated some sort of recording detection, although it was easy to circumvent. Therefore, if the attackers were able to record the activation command of the authorized owner it will be very difficult to implement a protection. However, such recording is difficult to obtain, while the approach of using any recorded talk to train an AI voice model is the real threat.

Smart personal assistants rely on voice biometrics to identify the authorized user; they use physiological patterns (physical shape of the vocal tract) and behavioral characteristics (how the person chooses to speak) [34]. The details of the specific algorithms implemented by the different SPAs are not public, but it is known that "voiceprint" use parameters such as pitch, timbre, intensity, accent of pronunciation or pronunciation habits (linguistic inflections), and cadence [35]. Our experimental results indicate that Amazon's Alexa employs a relatively permissive access control model, even suggesting training the system for a new voice just by saying "learn my voice". Furthermore, the system demonstrated a lack of robustness against AI-generated (synthetic) speech. Alexa consistently granted access to spoofed identities without detecting discrepancies from the enrolled biometric model. This suggests that the underlying recognition engine either utilizes a low-dimensional feature set for vocal verification or operates with a high tolerance threshold.

There is a very active area of research related to the detection of deepfake voice, that may help to address the attack presented in this paper. Initial systems for detection of fake voices were based on Machine Learning techniques [36] obtaining very good results for detecting voices generated using an imitation method based on signal processing tool. DeepSonar [37] uses deep neural networks for detecting

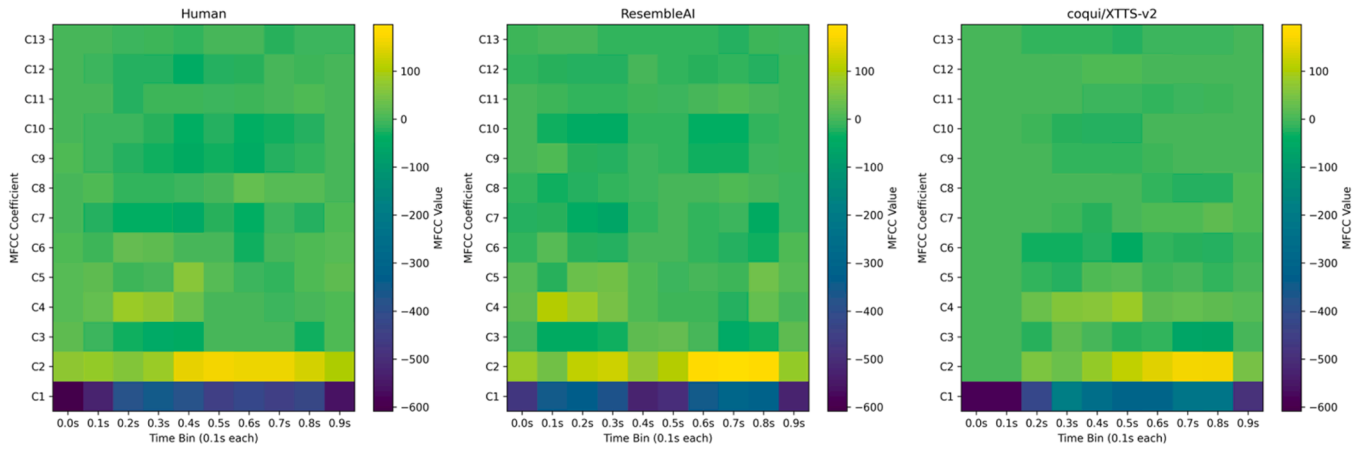

Fig. 10. Mel-frequency cepstral coefficients (MFCC) of the 3 most relevant audio signals.

AI-synthesized voices, through the introduction of a method based on identifying layer-wise neuron activation patterns. Deep4SNet [38] computes the histograms of the recordings to be used as the input of convolutional neural networks. And the most recent classifier found in the literature [39] uses Mel-Frequency Cepstral Coefficients (MFCC) of the recording for the analysis. All these methods claim accuracies over 90%, however when detection systems are trained with dataset, the models may learn the characteristics of the particular algorithm used to generate the fake audios and yield very good results as long as the fake audios are generated with the same technique.

In order to evaluate the possibility of distinguishing human voices from AI generated voices, we developed a website implemented as a game to engage anonymous users to classify voices as human or AI [40]. This game gathered the results of said users and served us to obtain conclusions on how difficult it is for human beings to identify AI generated voices.

The process for collecting evaluations is depicted in Fig. 11. When users access the website, they will first see an introduction explaining the purpose of the experiment. Then they will be directed to a form where they'll provide some personal information that may be relevant for future studies, such as whether they are native English speakers, their level of education, age, and country of origin. Once the form is completed, data are securely stored in the platform's database for later statistical analysis. On the next page, users will find a brief calibration stage to adjust their audio settings and get a sense of how the experiment works.

After clicking Play, the game starts. The system plays an audio recording that has been randomly selected from the database and asks the user to evaluate whether it is "Human", "Sounds Human", "Sounds AI", or "AI". With this information the system obtains the accuracy and also the certainty (if the person answered with confidence). As a result, the website shows the statistics of the current game (see Fig. 12), encouraging the user to "play" again, so we could get more results. Finally, the results stored in the database can be analyze using a Data Visualization tool.

The website used The Fake-or-Real Dataset created by the Lassonde School of Engineering of York University, Toronto, Canada [41], also available on Kaggle [42]. This dataset is a collection of >195,000 audios from real humans and computer-generated speech [43]. In addition, the website incorporated some of the audios that were used for testing the Smart Personal Assistants, so we were able to evaluate how difficult is for humans to distinguish between the human and AI cloned voices used to test the SPA's privacy levels. In total we collected 890 responses.

The results show that distinguishing between human and AI generated audios is indeed challenging. As shown in the Global Confusion

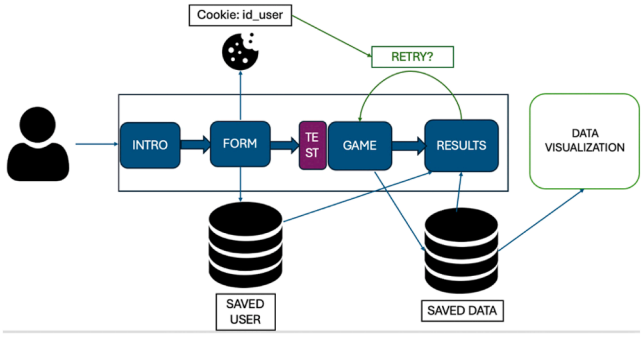

Fig. 11. Graphical description of the web application developed to obtain manual evaluation of Human and AI voices.

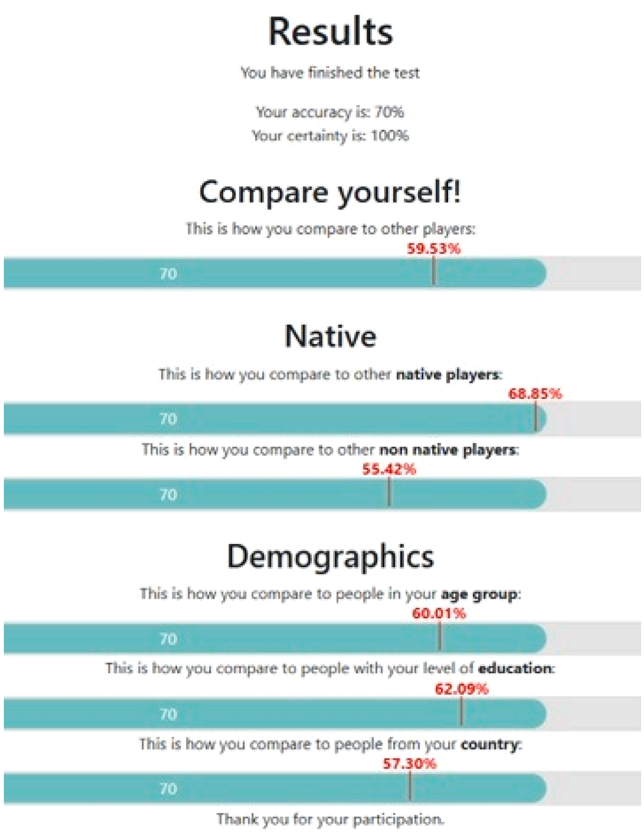

Fig. 12. Example of website results.

Matrix (Fig. 13), 280 AI audios and 228 human audios were correctly identified, while 169 human audios where mistakenly identified as AI and 213 AI generated audios were misidentified as Human. These values yield an accuracy of 57.1% (from 53.8% to 60.3% with a 95% confidence interval). Standard deviation is 1.66 and the p-value is less that 0.05. These results demonstrate the difficulty that humans face in identifying synthetic voices as non-human and also show how AI-generated speech models can effectively deceive humans. The reason for such low accuracy is because there are too many computer-generated audios misclassified as human (213 out of 493), even though 57% of the human audios where correctly classified as human. The system also collected demographic data, and it is interesting to observe that the highest accuracy is attained by the middle-age category, although the sample size is the smallest, so the uncertainty range is the largest. For the group younger that 30 years old the accuracy is 57.1% [51.3% – 62.8%], for ages between 30 and 50 the accuracy is 62.5% [51.6% –72.3%], and for ages greater that 50 the accuracy obtained is 56% [52.0% –60.4%]. All the results show a narrow margin above 50% which suggests that humans lack the psychoacoustic discernment necessary to identify sophisticated synthetic speech. This implies that a machine learning model to distinguish real and fake voices needs to outperform humans by detecting features invisible to the human ear. If voice cloning successfully replicates a subject's unique biometric signature but fails to simulate the organic nuances of human speech production, a dedicated Synthetic Speech Detection model will serve as a critical secondary authentication factor to improve security in SPAs.

Data obtained by the platform also highlights that synthetic voice generated with the most updated AI voice cloning tools are more likely to deceive humans, because the majority of the samples were classified as human. Fig. 14 shows the results corresponding to audios generated with three AI cloning systems: Lovo, PlayHT and ResembleAI. The graph shows the percentage of people who correctly identified these voices as AI. Lovo, which was discarded because the commands that it generated didn't sound natural or alike the owner's voice, obtained the worst results fooling humans. With a recognition rate of 33% as AI, it is a better algorithm than the synthetic voices in the database because only one-third-of the users classified these voices as AI, however the other AI cloning systems performed better because they fooled even more people. ResembleAI, which was the AI model able to unlock Apple HomePod, was only detected as AI voice by 27%. PlayHT was identified as AI only by 8% of the users, meaning that the sound is very human-like or very realistic.

The demographic data gathered before the user starts playing the

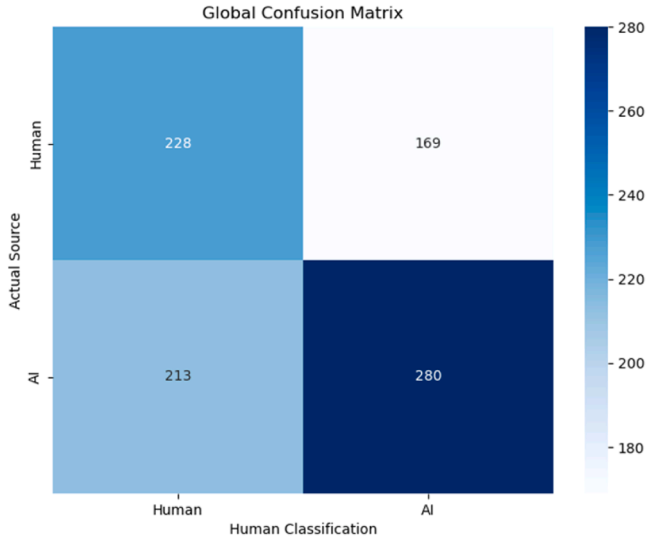

Fig. 13. Confusion Matrix of human classification of voices.

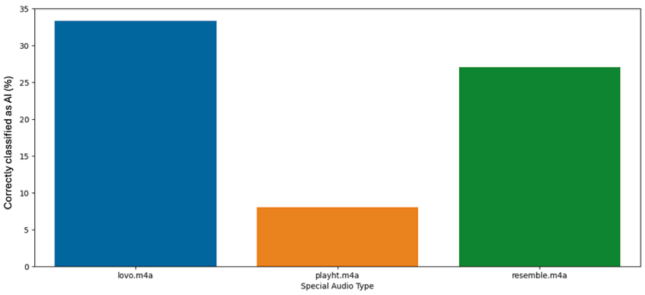

Fig. 14. Human classification of voices generated with AI models.

game may be very useful to carry out social science studies on such an emerging threat as vishing. As a matter of fact, the game can provide useful data to identify which population segments are more vulnerable to this kind of attack, so that effective policies can be developed to protect them. In addition, the game can be also used by more skilled users to keep improving their capabilities to distinguish between real and synthetic voices to avoid more sophisticated and high-impact attacks such as corporate espionage.

6. Conclusions

The tests conducted in this research have demonstrated that personal assistants from the most widespread brands do a good job distinguishing between actions that do not pose risks and commands that reveal personal information or perform specific actions. When harmless actions, such as playing music or checking the temperature, are requested, the devices accept any voice without the need for validation. However, when personal data is requested or actions such as sending messages or making calls are demanded, all devices require a user voice validation process. The tests carried out have shown that this validation is based on recognizing personal traits in the activation command ("Hey Siri," "Hey Google," or "Alexa").

The experimental tests presented in this article demonstrate that the use of synthetic voices poses a risk to accessing Personal Assistants. With text-to-speech systems, only harmless actions can be performed because the generated voice is generic and does not include the owner's characteristic traits. However, with recorded voices of the registered user, it has been possible to execute sensitive commands, even bypassing the attempted protections of Apple HomePod and Google Home Mini against recorded voices. Nevertheless, using a recorded activation command is not considered a significant threat since such recording is typically difficult to obtain.

The greatest success in attacking Personal Assistants has been achieved by applying generative AI to imitate the registered user's voice. The AI models have allowed for the generation of activation commands with the voice's characteristic traits of the device owner, enabling the execution of sensitive commands. The risk associated with this attack method is that the generative AI model is trained using the owner's voice recording that do not include any of the words that are part of the activation commands. The method PlayHT activated all the SPAs except the Apple HomePod. However, the current version can only be trained with specific sentences instead of free speech. On the other hand, Resemble.AI was able to surpass the validation algorithm of all the SPAs. In the case of the HomePod the authentication algorithm didn't achieve an internal confidence level as high as with the real owner's voice, but it was high enough to active the device. One main advantage of Resemble.AI is that it can be trained with any recording of the victim (a TED talk, a TV interview, a Conference, etc.). Finally, using SV2TTS, that can also be trained with any voice recording, the results were not that satisfactory. Sound quality was worse than with the other models, and only Alexa was vulnerable to the generated voice.

As a protection against this kind of attacks, it was proposed to

implement systems to differentiate human voiced from deepfake voices as part of the authentication algorithm [44]. A web-based system was developed to gather information about how humans classify synthetic voices and human voiced. It came as a surprise that humans are not very effective at this task, especially in the case of voices generated with the AI models analyzed. In the case of ResembleAI, which was the best impersonating the owner with all SPA devices, human users labeled these voices as human in most of the cases. Only 27% of the experiments correctly labeled these voices as synthetic (with a MoE Margin of Error of 12% for a confidence level of 90%). Therefore, the development of an automatic classifier to detect deepfake audios is a very challenging task.

CRediT authorship contribution statement

Clara Palacios-Castrillo: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Rafael Palacios:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Investigation, Formal analysis, Conceptualization. **Roberto Gesteira-Miñarro:** Writing – review & editing, Validation, Methodology, Investigation, Conceptualization. **Alejandro Chávez-Macías:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Gregorio López:** Writing – review & editing, Visualization, Validation, Resources, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Datasets used for evaluation of real voices and fake voices are public and have been cited

References

[1] Ruan S, Wobbrock JO, Liou K, Ng A, Landay JA. Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018;1(4):1–23. https://doi.org/10.1145/3161187.

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

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

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

[5] Valero C, et al. Analysis of security and data control in smart personal assistants from the user's perspective. Fut Gener Comput Syst 2023;144:12–23. https://doi.org/10.1016/J.FUTURE.2023.02.009.

[6] Fúster J, et al. Analysis of security and privacy issues in wearables for minors. Wirel Netw 2024;30(6):5437–53. https://doi.org/10.1007/S11276-022-03211-6.

[7] Edu JS, Such JM, Suarez-Tangil G. Smart home personal assistants. ACM Comput Surv (CSUR) 2021;53(6):1–36. https://doi.org/10.1145/3412383.

[8] Khanjani Z, Watson G, Janeja VP. Audio deepfakes: a survey. Front Big Data Jan. 2023;5:1001063. https://doi.org/10.3389/FDATA.2022.1001063/PDF.

[9] Schmitt M, Flechais I. Digital deception: generative artificial intelligence in social engineering and phishing. Artif Intell Rev 2024;57(12):1–23. https://doi.org/10.1007/S10462-024-10973-2.

[10] Siddiqi MA, Pak W, Siddiqi MA. A study on the psychology of social engineering-based cyberattacks and existing countermeasures. Appl Sci Jun. 2022;12(12):6042. https://doi.org/10.3390/APP12126042.

[11] Shaaban OA, Yildirim R, Alguttar AA. Audio deepfake approaches. IEEE Access 2023;11:132652–82. https://doi.org/10.1109/ACCESS.2023.3333866.

[12] Lorenzo-Trueba J, Fang F, Wang X, Echizen I, Yamagishi J, Kinnunen T. Can we steal your vocal identity from the internet?: initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data. Speak Lang Recognit Workshop ODYSSEY Mar. 2018:240–7. https://doi.org/10.21437/Odyssey.2018-34.

[13] Luong HT, Yamagishi J. NAUTILUS: a versatile voice cloning system. IEEE/ACM Trans Audio Speech Lang Process 2020;28:2967–81. https://doi.org/10.1109/TASLP.2020.3034994.

[14] Bilika D, Michopoulou N, Alepis E, Patsakis C. Hello me, meet the real me: audio deepfake attacks on voice assistants. Accessed, https://arxiv.org/abs/2302.10328v1; 2025.

[15] Wang Y, et al. Tacotron: towards end-to-end speech synthesis. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH; Mar. 2017. p. 4006–10. https://doi.org/10.21437/Interspeech.2017-1452.

[16] Le M, et al. Voicebox: text-guided multilingual universal speech generation at scale. Adv Neural Inf Process Syst 2025. Accessed[Online]. Available, https://proceedings.neurips.cc/paper_files/paper/2023/hash/2d8911db9ecedf866015091b28946e15-Abstract-Conference.html.

[17] C. Jemine, "Master thesis : automatic Multispeaker voice cloning." [Online]. Available: http://lib.uliege.be.

[18] Voice Cloning Studio - a Hugging Face Space by hasanbasbunar. Accessed: Nov. 16, https://huggingface.co/spaces/hasanbasbunar/Voice-Cloning-XTTS-v2; 2025.

[19] Noun_Project. Noun project: free icons & stock photos for everything. Accessed: Mar. 07, https://thenounproject.com/; 2025.

[20] Buck M. Audio quality: performance testing of information appliances. Audio Eng Soc Mar. 01, 2001.

[21] AI Voice Generator: Realistic Text Speech AI Voiceover | Play. Accessed, https://play.ht/; Mar. 15, 2024.

[22] AI Voice Generator with Text to Speech and Speech to Speech. Accessed, https://www.resemble.ai/; Mar. 15, 2024.

[23] AI Voice Generator: Realistic text to speech & voice cloning. Accessed, https://lovo.ai/; Mar. 15, 2024.

[24] AI Voice Generator, Text to speech, #1 best AI voice. Accessed, https://speechify.com/; Mar. 15, 2024.

[25] AI Voice Cloning Online: Clone your voice in seconds. Accessed, https://murf.ai/voice-cloning; Mar. 15, 2024.

[26] AI Voice Generator & Text to Speech | ElevenLabs. Accessed, https://elevenlabs.io/; Mar. 15, 2024.

[27] Elai.io - AI Voice Cloning: Clone Your Voice Effortlessly. Accessed, https://elai.io/voice-cloning; Mar. 15, 2024.

[28] Shen J, et al. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. Institute of Electrical and Electronics Engineers Inc.; Sep. 2018. p. 4779–83. https://doi.org/10.1109/ICASSP.2018.8461368.

[29] Mandeel AR, Al-Radhi MS, Csapó TG. Speaker adaptation experiments with limited data for end-to-end text-to-speech synthesis using Tacotron2. Infocommunications J. 2022;14(3):55–62. https://doi.org/10.36244/ICJ.2022.3.7.

[30] Jemine C. GitHub - CorentinJ/real-time-voice-cloning: clone a voice in 5 s to generate arbitrary speech in real-time. Accessed: Sep. 07, https://github.com/CorentinJ/Real-Time-Voice-Cloning; 2025.

[31] Casanova E, et al. XTTS: a massively multilingual zero-shot text-to-speech model. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH; Jun. 2024. p. 4978–82. https://doi.org/10.21437/Interspeech.2024-2016.

[32] Li K, Lu X, Akagi M, Unoki M. Contributions of Jitter and Shimmer in the voice for fake audio detection. IEEE Access 2023;11:84689–98. https://doi.org/10.1109/ACCESS.2023.3301616.

[33] Warren K, et al. Pitch imperfect: detecting audio deepfakes through acoustic prosodic analysis. Accessed: Nov. 18, 2025. [Online]. Available, https://arxiv.org/pdf/2502.14726; Feb. 2025.

[34] Vasan A. Voice biometrics: a detailed walkthrough. Accessed, https://www.parloa.com/knowledge-hub/voice-biometrics/; Mar. 03, 2026.

[35] "How do voice recognition biometrics work?". Accessed, https://stytch.com/blog/what-is-voice-biometric-authentication/; Mar. 03, 2026.

[36] Rodríguez-Ortega Y, Ballesteros DM, Renza D. A machine learning model to detect fake voice. Commun Comput Inf Sci 2020;1277:3–13. https://doi.org/10.1007/978-3-030-61702-8_1/TABLES/4.

[37] Wang R, et al. DeepSonar: towards effective and robust detection of AI-synthesized fake voices. In: MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia; Oct. 2020. p. 1207–16. https://doi.org/10.1145/3394171.3413716/SUPPL_FILE/3394171.3413716.MP4.

[38] Ballesteros DM, Rodriguez-Ortega Y, Renza D, Arce G. Deep4SNet: deep learning for fake speech classification. Expert Syst, Appl 2021;184:115465. https://doi.org/10.1016/J.ESWA.2021.115465.

[39] Tanveer MU, Munir K, Amjad M, Rehman AU, Bermak A. Unmasking the fake: machine learning approach for deepfake voice detection. IEEE Access 2024;12:197442–53. https://doi.org/10.1109/ACCESS.2024.3521026.

[40] Chávez A, López G, Palacios R. Human or AI?. Accessed, https://human-or-ai.icai.comillas.edu/; Feb. 03, 2025.

[41] Datasets – APTLY and LaSSoftE. Accessed, https://bil.eecs.yorku.ca/datasets/; Feb. 03, 2025.

[42] The Fake-or-Real (FoR) Dataset (deepfake audio). Accessed, https://www.kaggle.com/datasets/mohammedabdeldayem/the-fake-or-real-dataset; Feb. 03, 2025.

[43] Reimao R, Tzerpos V. FoR: a dataset for synthetic speech detection. In: 2019 10th International Conference on Speech Technology and Human-Computer Dialogue; Oct. 2019. https://doi.org/10.1109/SPED.2019.8906599. *SpeD 2019*.

[44] Latorre JM, Cerisola S, Ramos A, Palacios R. Analysis of stochastic problem decomposition algorithms in computational grids. Ann Oper Res 2009;166(1):355–73. https://doi.org/10.1007/s10479-008-0476-1.

Clara Palacios-Castrillo is a Telecommunications Engineering student (Class of 2025) at the ICAI School of Engineering, Comillas Pontifical University in Madrid, where she also pursues a minor in Biomedical Engineering. For the 2024-25 academic year, she is an exchange student within the Department of Electrical and Computer Engineering at Carnegie Mellon University (CMU). Clara is currently a research student at both the Institute for Research in Technology (Comillas) and the Robotics Institute at CMU.

Rafael Palacios received the B.S., M.S. and PhD degrees from the ICAI School of Engineering, Comillas Pontifical University, Madrid, Spain. He joined the ICAI School of Engineering, as an Assistant Professor, and the Institute for Research in Technology, as a Researcher, in 1998, obtained Tenure in 2004, and became a Full Professor in 2020. He has been the Head of the Programs in telecommunications engineering and computer science from 2012 to 2024, and coordinator of the Master in Cybersecurity from 2019 to 2021. He is currently a Visiting Professor with Cybersecurity at MIT Sloan - CAMS at the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA.

Roberto Gesteira-Miñarro received the B.S. degree in telecommunications engineering and the M.S. degree in telecommunications engineering and cybersecurity from the ICAI School of Engineering, Comillas Pontifical University, Madrid, Spain, in 2020 and 2022, respectively. He is currently pursuing the Ph.D. degree in vehicle cybersecurity with the IIT. He is also studying a B.S. degree in mathematics from the Universidad Nacional de Educaci..n a Distancia (UNED), Madrid. He developed his master thesis at the Institute for Research in Technology (IIT) regarding vehicle cybersecurity. Mr. Gesteira-Mi..arro received the Extraordinary Award from ICAI in 2020 and the Best Final Degree Project Award from COITT/AEGITT in 2021. He achieved two Honorific Mentions, two Bronze Medals, and one Silver Medal at the International Mathematics Competition (IMC). He also enjoys learning hacking and exploitation techniques in CTF platforms.

Alejandro Chávez-Macías holds BSc. and MSc. Degrees in Telecommunications Engineering and a BSc. in Business Administration from Comillas Pontifical University (ICAI-ICADE), Madrid, SPAIN. For his Master's Thesis, he engineered a cloud-native platform using Azure and Node.js to facilitate large-scale research into AI-generated audio classification. Currently, Alejandro works as a Data Engineer, specializing in cloud-based architectures, ETL processes, and large-scale data processing. His expertise spans real-time and batch processing systems, with a core focus on automating data workflows and enhancing system reliability within enterprise environments.

Gregorio López received the Ph.D. degree in telecommunications engineering from the Universidad Carlos III de Madrid (UC3M), in 2014. He is currently an Associate Professor with the ICAI School of Engineering, Comillas Pontifical University, where he is also the Coordinator of the M.S. in cybersecurity and a Senior Researcher with the Institute for Research in Technology. He has gained wide experience in close-to-market research gained through his participation in more than ten national and European research projects. As a result of his research activity, he holds a European patent and has published more than 50 papers in top-tier conferences and journals, receiving more than 1500 citations. His current research interests include cybersecurity in the IoT/OT, AI for cybersecurity and cybersecurity in AI, human and technological factors in cybersecurity, and children and adolescents online, having been the Coordinator of European H2020 Project RAYUELA (empoweRing and educAting YoUng pEople for the Internet by pLAying), which addresses this latter topic.