Computers & Security 134 (2023) 103448

Contents lists available at ScienceDirect

Computers & Security

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

Security and privacy problems in voice assistant applications: A survey

Jingjin Li [a,*], Chao Chen [b], Mostafa Rahimi Azghadi [a], Hossein Ghodosi [a], Lei Pan [c], Jun Zhang [d]

[a] College of Science and Engineering, James Cook University, Douglas, 4811, QLD, Australia
[b] School of Accounting, Information System and Supply Chain, RMIT University, Melbourne, 3000, VIC, Australia
[c] College of Science and Engineering, Deakin University, Geelong, 3220, VIC, Australia
[d] College of Science and Engineering, Swinburne University of Technology, Hawthorn, 3122, VIC, Australia

ARTICLE INFO

Keywords:
ASR
Security
Privacy
SI
Voice assistant
Attack
Defense

ABSTRACT

Voice assistant applications have become omniscient nowadays. Two models that provide the two most important functions for real-life applications (i.e., Google Home, Amazon Alexa, Siri, etc.) are Automatic Speech Recognition (ASR) models and Speaker Identification (SI) models. According to recent studies, security and privacy threats have also emerged with the rapid development of the Internet of Things (IoT). The security issues researched include attack techniques toward machine learning models and other hardware components widely used in voice assistant applications. The privacy issues include technical-wise information stealing and policy-wise privacy breaches. The voice assistant application takes a steadily growing market share every year, but their privacy and security issues never stopped causing huge economic losses and endangering users' personal sensitive information. Thus, it is important to have a comprehensive survey to outline the categorization of the current research regarding the security and privacy problems of voice assistant applications. This paper concludes and assesses five kinds of security attacks and three types of privacy threats in the papers published in the top-tier conferences of cyber security and voice domain.

1. Introduction

Naturally, people communicate through voice. Three technologies have been proposed to facilitate human-computer interaction through voice, including automated speech recognition (ASR), natural language processing (NLP), and speech synthesis (SS). NLP enables machines to comprehend human intents, and SS enables machines to talk. The research of voice assistant applications started in the 1950s with continuous refinement. Fig. 1 shows the progress of voice assistant applications over time. The Hidden Markov Model (HMM) method, a statistical model-based approach, has increasingly taken the lead in voice recognition research since the 1980s. As voice recognition technology thrives with the more advanced deep learning algorithm, it integrates with more and more devices. In 2011 when the iPhone 4S was released, the world's first mobile phone personal voice assistant Siri was known and opened a new chapter for voice assistant applications.

Nowadays, voice assistant applications prosper with a large portion of the market. The voice assistant has brought huge convenience to our daily life and greatly changed how humans interact with computers. Almost every smart device has a built-in voice assistant. Because of the development and wildly use of voice assistant applications, privacy and security problems emerge. Often, users may not even notice their conversation has been recorded or mistakenly awoken by the voice assistant. Taking Siri as an example, several times it heard other people saying "Are you serious?" or "a series of...", it just started to voice recognition and write down what it heard on my screen. A voice assistant is easily activated by accident, which malicious attackers could exploit. Plausible but severe threats include bank transfers, buying virtual products, fabricating messages to your close friends or families asking for money, stealing your credential information, and many others. Large financial and emotional losses may occur if voice assistant applications are breached. Thus, the security and privacy problems within voice assistant applications should be followed with interest. Users become more focused on taking complete control of their voice assistant applications, knowing the potential attack and defense methods.

* Corresponding author.
E-mail addresses: ginger.li@my.jcu.edu.au (J. Li), chao.chen@rmit.edu.au (C. Chen), mostafa.rahimiazghadi@jcu.edu.au (M. Rahimi Azghadi), hossein.ghodosi@jcu.edu.au (H. Ghodosi), l.pan@deakin.edu.au (L. Pan), junzhang@swin.edu.au (J. Zhang).

https://doi.org/10.1016/j.cose.2023.103448
Received 12 March 2023; Received in revised form 19 July 2023; Accepted 21 August 2023
Available online 25 August 2023

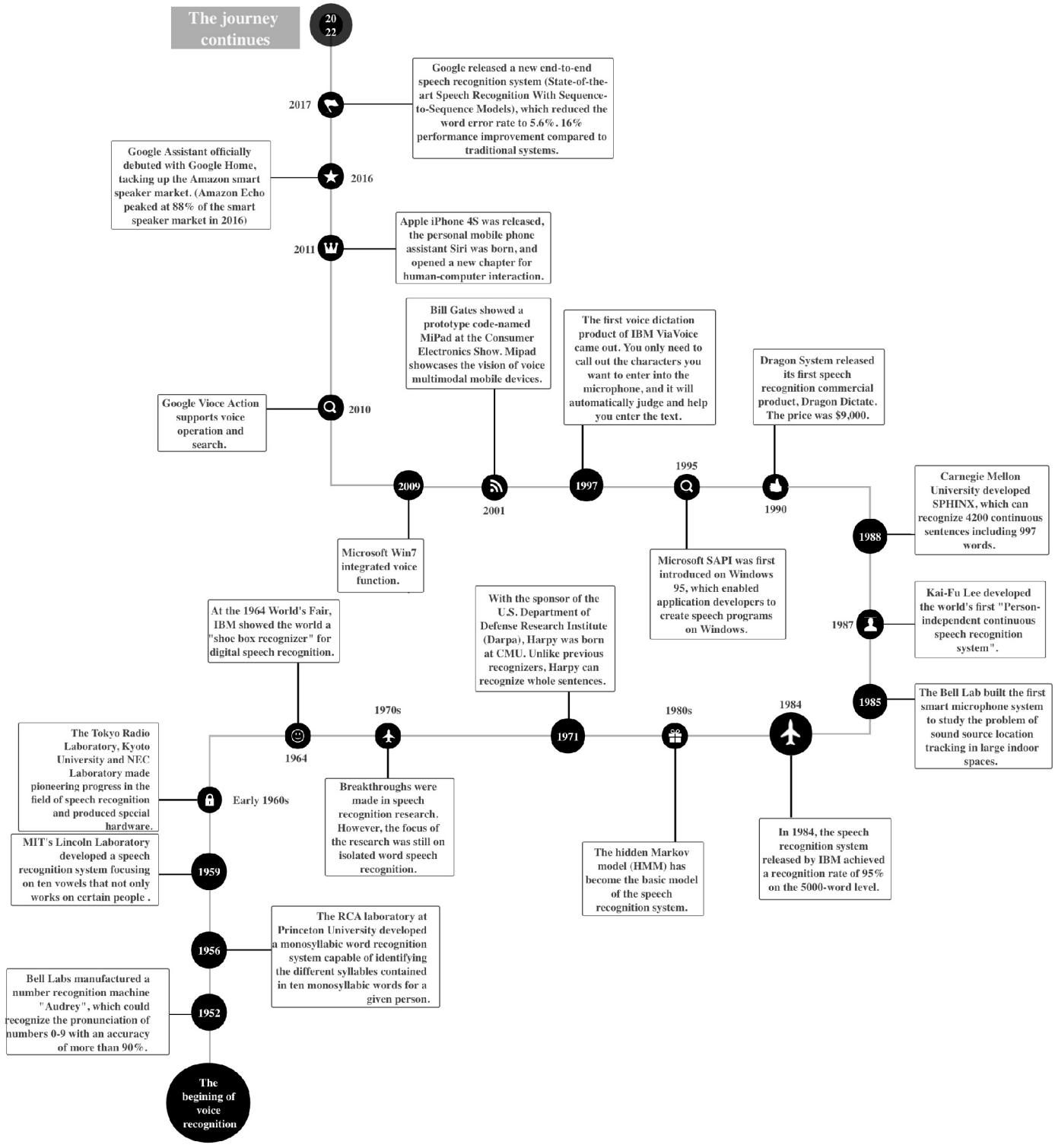

Fig. 1. The timeline that outlined the development and progress of voice assistant applications.

With many papers published in recent years about new attack techniques and defense means, a comprehensive outline of their development is needed. However, there are two surveys in the voice assistant security and privacy domain. Cheng et al. Cheng and Roedig (2022) focused on security and privacy problems that voice assistants have in using acoustic channels. Acoustic channel attacks manipulate audio signals to deceive or compromise voice assistants. Four main problems were included in Cheng et al. Cheng and Roedig (2022). They are access control loss, acoustic DoS attack, voice privacy loss, and malicious use of acoustic sensing. Compared to their survey, this survey also covers none acoustic channel attacks. Non-acoustic channel attacks exploit non-acoustic communication channels to gain unauthorized access or extract sensitive information from voice assistants. Also, their survey only covers research before the end of 2020, and this survey covers more recent research published in top security conferences until 2022. In this survey, during 2020-2022, quite a few papers about new side-channel attacks were researched and had impressive results. Another difference is that this survey divided the attacks towards ASR and SI

models. Because ASR and SI are the two major functions that voice assistant applications have, some applications have both, and some have only one. It is useful for users to know which function their applications have and what type of threats correspond to each application.

The other survey by Yan et al. Yan et al. (2022) covers almost the same security and privacy problems mentioned in Cheng et al. Cheng and Roedig (2022). However, it also does not cover any research after 2020. Also, Yan et al. Yan et al. (2022) categorize defensive methods from a system designer's perspective, which differs from this survey and Cheng et al. Cheng and Roedig (2022). In this survey, defensive methods are introduced regarding specific attacks. Through this survey, some defensive methods are effective in multiple types of attacks, and some mitigate one type of attack but are prone to another type of attack, which helps the users or producers have more comprehensive information when choosing the defensive methods.

This survey aims to make a comprehensive and clear outline of the security and privacy issues of voice assistant applications. The papers included in this survey are from the top four cyber security conferences and Interspeech, a conference focusing on the speech domain. The aspects that are included are as follows:

1. Technical attacks that were targeted towards ASR models and SI models. Including machine learning attacks that targeted the software, frequency modulation that exploits the hardware, malicious skills hidden in the third-party market and policy loopholes that were not refined quickly enough to catch up on the development of voice assistants.
2. Defensive methods have been researched and proven effective in ASR and SI models. Usually, the defensive means can be divided into detection and prevention. Some methods may provide both means.
3. We have security and privacy issues when using voice assistant applications beyond technical threats. With more and more younger users, third-party regulations and policies should be refined.

Contributions. This paper provides a comprehensive summary of technical attacks with impressive experiment results and feasible defensive methods corresponding to each attack. Nontechnical threats in the voice assistant application market are included to safeguard the user. The contributions are concluded as follows:

- From a user's standpoint, this survey is, as far as we are aware, the most thorough investigation of voice assistant application security. Our study includes both market policy issues and technology risks. We provide a comprehensive overview of the state of the art, development, major difficulties, and future prospects for voice assistant application security research based on a thorough literature review of pertinent attacks and countermeasures.
- We classify pertinent assaults by attack techniques and structure the attack literature according to the voice assistant's systems. In order to properly identify, comprehend, and analyze the security risks against voice assistants, the organization assists in bridging the gap between a large category of seemingly unrelated attacks and vulnerabilities.
- To systematize the countermeasures against various attacks, we base them on defensive tactics. We present a qualitative evaluation of existing solutions by the installation cost if the defense requires additional devices, usability, and security and make useful recommendations in order to help users select protection based on the type of danger they may encounter.

The remaining portions of this essay are structured as follows: The introduction to voice assistant applications in Section 2 is brief. The taxonomy of assaults on ASR and SI models as well as the taxonomy of countermeasures that may be applied to ASR and SI models are also introduced in Section 2. The attacks that take advantage of the voice assistant's ASR function's weaknesses are described in detail in Section 3 along with the corresponding defenses. The attacks against SI models are described in detail in Section 4, along with systematized defense tactics that can stop or at least slow them down. The security and privacy issues outside of technological assaults are summarised in Section 5. In Section 6, we go through issues with the current research and potential future approaches for voice assistant applications. The survey is concluded in Section 7.

2. Preliminaries of voice assistant applications

This section provides background information on voice assistant apps, including a definition of key terms, a list of categories, and a description of the process for each type of voice assistant application.

2.1. Voice assistant components and speech recognition workflow

There are two kinds of voice assistant models — automatic speech recognition (ASR) and speaker identification (SI). As shown in Fig. 2 and Fig. 3, the first step in creating a voice recognition model is translating the spoken language into text. Speech recognition is much more challenging to solve than machine translation. A machine translation system's input is usually printed text that differentiates between individual words and word strings. The voice input used by a speech recognition system is far more complicated than written text and spoken language, especially with ambiguity. When two people communicate, they frequently infer the term in the conversation in the context and often read a lot of latent information from the tone, facial expressions, and gestures the other party uses. The speaker regularly rectifies what has been said and repeats important material by rephrasing. It is challenging to train an automated system to detect and comprehend speech. To provide a compact digital representation of the sound wave, each sampled value is quantized throughout the speech recognition process. A feature vector characterizing the spectral content is retrieved for each frame from which the sampled values are situated in overlapping frames. The words that the speech represents are identified based on the features of the voice signal. The five steps that make up the voice recognition process are described as follows:

Step 1. Voice Signal Acquisition

Voice signal acquisition is the foundation of voice signal processing. A voice signal acquisition system typically receives inputs through a microphone. Subsequently, the sound wave is transformed from a voltage signal by the microphone to a digital signal handled by an A/D device like a sound card. Voice signal acquisition and processing systems based on single-chip microcomputers and DSP chips are utilized extensively for unfavorable on-site conditions, limited space, and numerous specific equipment. The essential hardware for voice assistant apps includes sound cards, speakers, microphones, and many alike. Sound cards are crucial to process voice signals through signal filtering, amplification, A/D conversion, and D/A conversion. Modern recording software tools activate the sound card to harvest voice signals as voice recordings.

Step 2. Speech Signal Pre-processing

After collecting the speech signal, pre-processing operations must be completed, including filtering, A/D conversion, pre-emphasis, and endpoint detection. Filtering primarily serves the two goals of preventing aliasing interference and suppressing the 50 Hz power frequency interference. The voice analog signal is converted into a digital signal via A/D conversion. The signal is quantized during A/D conversion, and the quantization error, also known as quantization noise, is the difference between the quantized signal value and the original signal value. Pre-emphasis processing aims to improve the signal's high-frequency content, flatten its spectrum, and maintain its full frequency range from low to high frequency. Endpoint detection involves extracting the beginning and conclusion of speech from a speech-containing signal. Effective endpoint identification removes background noise in silent periods. Two

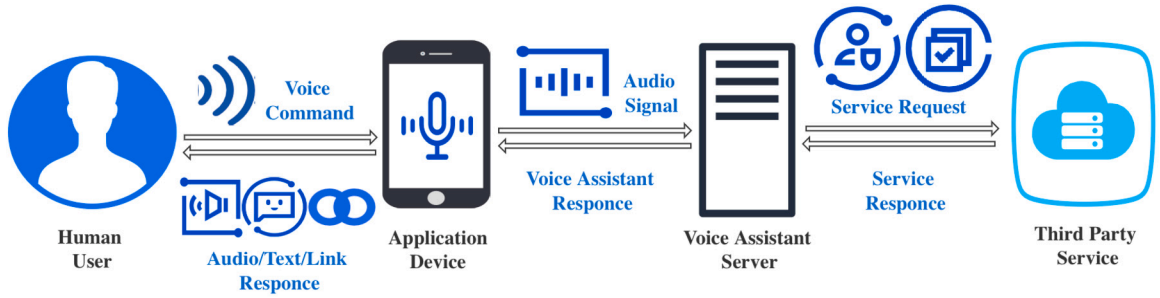

Fig. 2. Workflow of voice assistant application service.

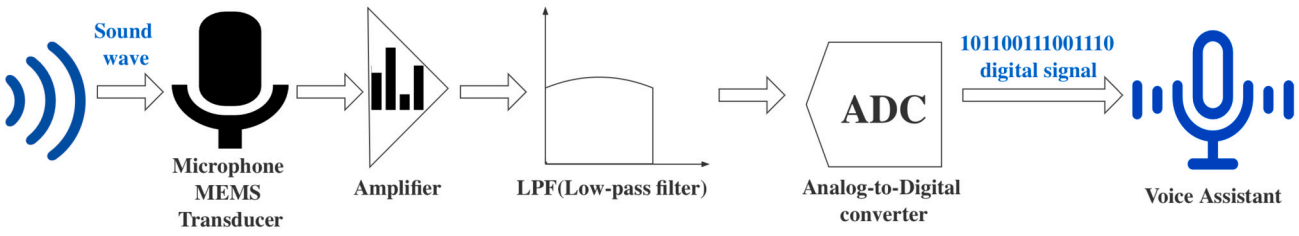

Fig. 3. Microphone components and voice signal capturing and pre-processing workflow.

popular approaches work on different features — time-domain features and frequency-domain features. The time domain feature approach uses the voice volume and zero-crossing rate to identify endpoints with the advantage of a minimal amount of calculation. However, the time domain feature approach often leads to incorrect evaluation of air sounds, and differing volume calculations will also result in varied detection outcomes. On the other hand, the frequency domain feature technique uses variations in the sound spectrum and entropy detection to identify speech at a high computation cost.

Step 3. Feature Parameter Extraction of the Speech Signal

The frequency of human speech is below 10 kHz. Shannon's sampling theorem requires the sampling frequency to be at least twice the maximum speech frequency present in the speech signal. The signal is often broken into blocks (also known as frames). Frames should overlap each other to prevent losing crucial information. Microphones collect waveforms of sound. It is important to extract distinctive information to separate the words from the collected data. Techniques for linear predictive coding are frequently employed to extract voice components. The fundamental tenet of linear predictive coding is that speech signal sampling points are correlated so that a linear combination of numerous previous sampling points helps predict the values of the present and subsequent sampling points. The linear prediction coefficient is calculated to reduce the mean square error between the anticipated and actual values.

Step 4. Vectorization

Vector quantization (VQ) is a data compression and coding method. In scalar quantization, a dynamic range is split into several subintervals, where each sub-interval has a representation value. This representative value is used to determine the value for an input scalar signal that falls inside the sub-interval during quantization. Due to scalar quantization, the semaphore is a one-dimensional scalar. VQ transforms a scalar into a one-dimensional vector from the perspective of linear space to quantify the vector. VQ separates the vector space into numerous little sections. A representative vector replaces the vectors in the section during quantization for each small section. VQ integrates various scalar values into a vector (or feature vector generated from a frame of speech data) to provide overall quantization in multi-dimensional space and enable data compression with minimal information loss. In a hidden Markov model, the input observation symbol can alternatively be the vector quantized feature vector.

Step 5. Speech Recognition

A typical speech recognition task is recognizing words and phrases because words are sequences of letters. A recognition system receives feature parameters from the speech signal as the input, like the LPC predictive coding parameters. Using Bayesian decision-making with maximum likelihood, three typical approaches are used in speech recognition: template matching, stochastic model, and probabilistic parsing.

- In template matching, a template is generated and stored while a user pronounces each phrase during the training stage. Each template in the template library is a feature vector. During the recognition stage, the input speech's feature vector sequence is iteratively compared to each template in the template library for the best match.
- The hidden Markov model (HMM) is the most popular method among stochastic models. HMM is a time-varying process that transits from one reasonably stable feature to another characteristic. With adequate time, the speech signal's properties gradually stabilize.
- Probabilistic parsing is used for continuous voice recognition across broad length ranges. While individuals speak the same phonetics, significant differences in the corresponding spectrograms and their modifications exist among individuals.

Last but not least, several other voice recognition techniques exist especially artificial neural network-based approaches for voice recognition, including the BP neural network, the Kohcmen feature mapping neural network and other networks with deep learning (Fig. 3).

2.2. Speaker identification workflow

The Speaker Identification (SI) system is often referred to as speaker recognition. SI consists of two stages: speaker identification and speaker confirmation. Speaker identification is a one-to-one mapping, and speaker confirmation is a many-to-one mapping. SI determines whether multiple speakers are present in a record and validates a speaker's identity by analyzing and processing the speech signal of the speaker. SI creates a reference template or model by extracting unique characteristics from the original voice signal before recognizing a speaker according to the predetermined criteria. In a SI task, the system extracts the speaker's personality traits by averaging the semantic information in

the speech signal, emphasizing the individual's distinctive characteristics; in a speech recognition task, the system normalizes the differences between different people's speeches as much as possible. The waveform of the speaker's speech reflects differences in pronunciation organs and habits, revealing each person's speech as a distinct personal trait that serves as an objective assurance of the speaker's identity.

Depending on the speaker numbers, speaker identification has two categories: closed set and open set. A closed set SI requires reflecting the number of speakers in the set to a closed set; on the contrary, an open set SI requires disregarding the number of speakers. Only a comparison and judgment between a reference model and the test speech are required for validation.

Speaker identification may be broken down into three groups: text-related, text-independent, and text-prompted. The speaker's pronunciation of essential words and phrases is used as a training text by text-related SI, and the same information is uttered during recognition. The recognition object is a free speech signal, and the text-independent speaker identification technique does not define speech content during training or recognition.

The training stage and the recognition stage are the two key phases. A template or model of each speaker is created during the training phase using feature extraction and the training corpus for each speaker in the speaker set. The speech to be recognized is broken down into its component characteristics at the recognition stage and compared to the template or model created during the system training. In speaker identification, the recognition outcome is the speaker corresponding to the model with the highest predicted speech similarity. Decide speaker confirmation by determining if the similarity between the test tone and the claimed speaker's model is higher than a predetermined threshold. The following fundamental issues affect the SI system's realization:

1. Preprocessing speech signals and feature extraction or extracting parameters can describe speaker characteristics.
2. Establishing the speaker model and establishing the model's training parameters.
3. Calculating the test speech's similarity to the speaker model.
4. Identification and technique for choosing. Confirmation or identification of the speaker.

Three categories can be used to implement SI:

1. Template matching — A reference template is a set of feature vectors to characterize the sequence of feature vectors. During the training process, feature vectors are extracted from the training sentences of each speaker to extract the feature vector sequence. During identification, a subject's template is compared with each reference template. The outcome of matching is frequently the accumulated distance between the feature vectors, measured as part of the matching process. VQ and dynamic time normalization (DTW) are template-matching techniques most often utilized.
2. Probabilistic model — An effective feature vector from pronunciations accurately characterizes the speaker's feature vector's distribution in the feature space. A mathematical model is constructed using statistical characteristics. A few model parameters serve to represent and store mathematical models. The feature vector of the test speech is compared to the mathematical model used to describe the speaker. The similarity between the test speech and the model is computed and helps make the recognition decision. The most widely used model is HMM because HMM correctly captures the properties of human vocal tract alterations and provides a reasonable description of stationarity and variability.
3. Artificial neural networks (ANN) — ANN is self-organized and self-learning, which may enhance its performance over time. ANN's features may be utilized to effectively extract speakers' personality traits from audio samples to implement SI systems.

Several performance metrics for evaluating the SI system include recognition rate, training duration, number of training corpora, reaction time, speaker set size, speaking mode, pricing, and number of training corpora. There are several assessment indicators for various events. The recognition rate is the most crucial factor, and it must be assured first to serve as the baseline for all other performance measures. Accurate and false recognition rates are frequently employed in voice recognition systems. The speaker confirmation mechanism determines the erroneous rejection and acceptance rates. The two are at odds with one another. Different sizes are needed for various events. The equal error probability is crucial for assessing speaker confirmation since it states that the two are equivalent if a particular judgment threshold is met.

2.3. Metrics

The metrics commonly used to evaluate the performance of voice assistants in converting spoken words into text are the Word Error Rate (WER) and Sentence Error Rate (SER). WER measures the proportion of words in the recognized text that differ from the words in the reference (correct) transcript. It is calculated by dividing the number of added, changed, or removed words by the total number of words in the reference transcript.

On the other hand, SER measures the number of sentence recognition errors, such as incorrect or missing sentences, divided by the total number of sentences in the reference transcript.

It is worth noting that SER is typically 2 to 3 times higher than WER due to the cumulative effect of errors within sentences. However, despite the higher error rate, SER is often neglected in evaluations, and more emphasis is placed on WER. This is because WER provides a more granular analysis of individual word errors, which is crucial for assessing the accuracy of voice assistants' transcription capabilities.

In addition to WER and SER, another important metric to consider is the Attack Success Rate (ASR). ASR measures the effectiveness of attacks against voice assistants by evaluating the proportion of successful adversarial attempts in manipulating or deceiving the system's speech recognition. ASR reflects the vulnerability of voice assistants to various adversarial techniques and is an essential metric in assessing the security and robustness of these systems.

2.4. Taxonomy

We develop a taxonomy of security and privacy problems in the voice assistant domain to define the attacks against voice assistants. This taxonomy classifies voice assistants' security and privacy risks recently published. Our taxonomy investigates various target models, adversarial information, and attack strategies. We further classify publications in the target model level category based on probabilistic models and target machine learning model types, such as DNN, RNN, CNN, and many alike. Popular apps are also used in various empirical studies, including Amazon Alexa, Google Assistant, and Microsoft Cortana. We categorize articles according to adversarial knowledge levels as black-box, grey-box, and white-box attacks on the target model. The taxonomy of attacks that threaten voice assistant models is shown in Fig. 4. The current attacks on SI models include spoofing, backdoor, adversarial, and hidden command attacks. Attacks in ASR models include dolphin attacks, adversarial attacks, and hidden command attacks.

We also developed a taxonomy for defensive methods in the voice assistant domain. The taxonomy of defensive methods is shown in Fig. 5. The defensive methods were listed based on different attacks that they mitigate. Furthermore, each mitigation method was categorized based on its mitigation types: detection and prevention. The defensive methods also were categorized based on whether they needed extra devices when they are deployed to protect the voice assistant applications.

We categorize all published publications that discuss attacks on ASR and SI models and then choose a few exemplary studies to put in Tables 1 and 3. Each chosen paper either revealed new ways to exploit

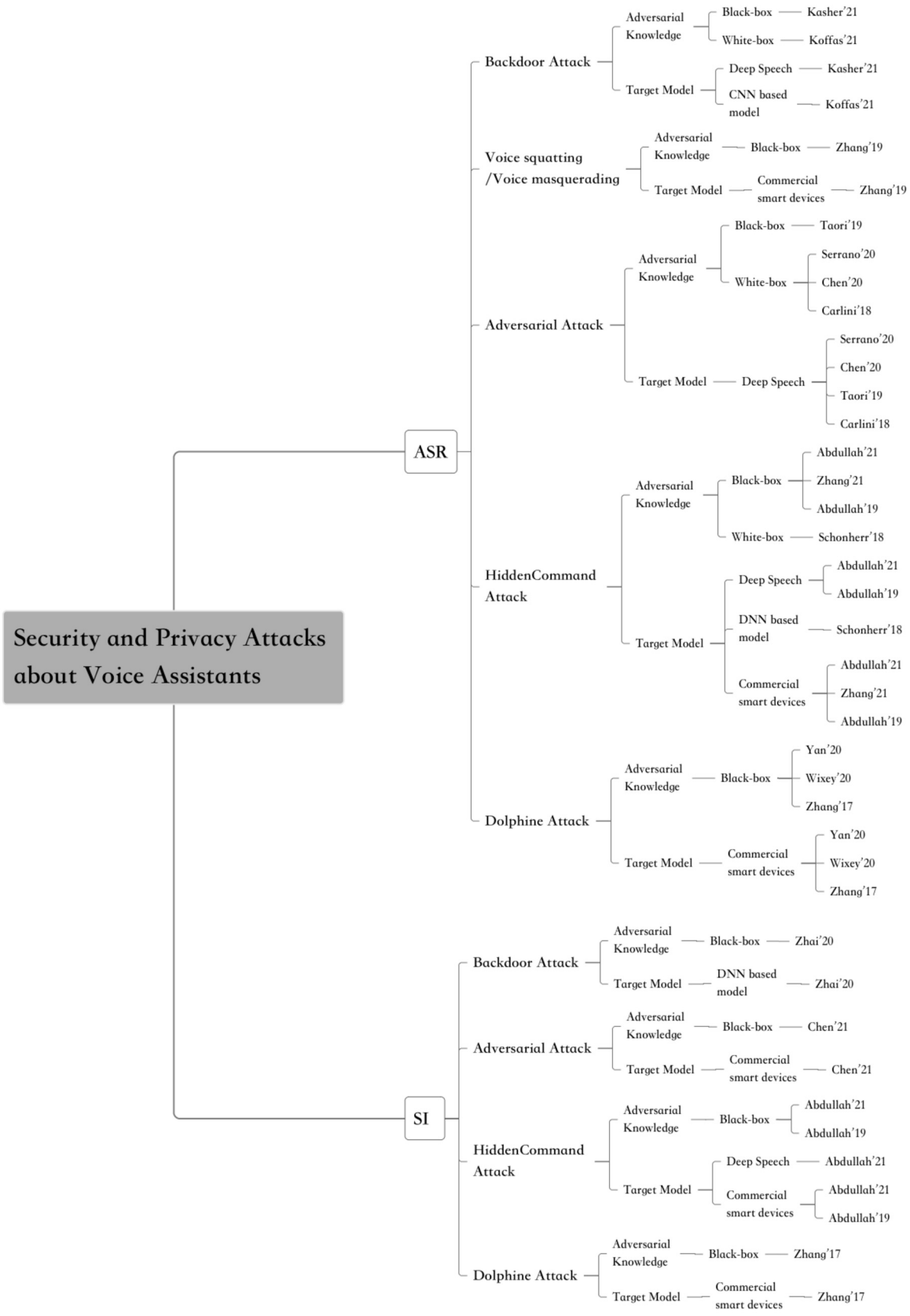

Fig. 4. Taxonomy of security and privacy attacks towards voice assistant applications.

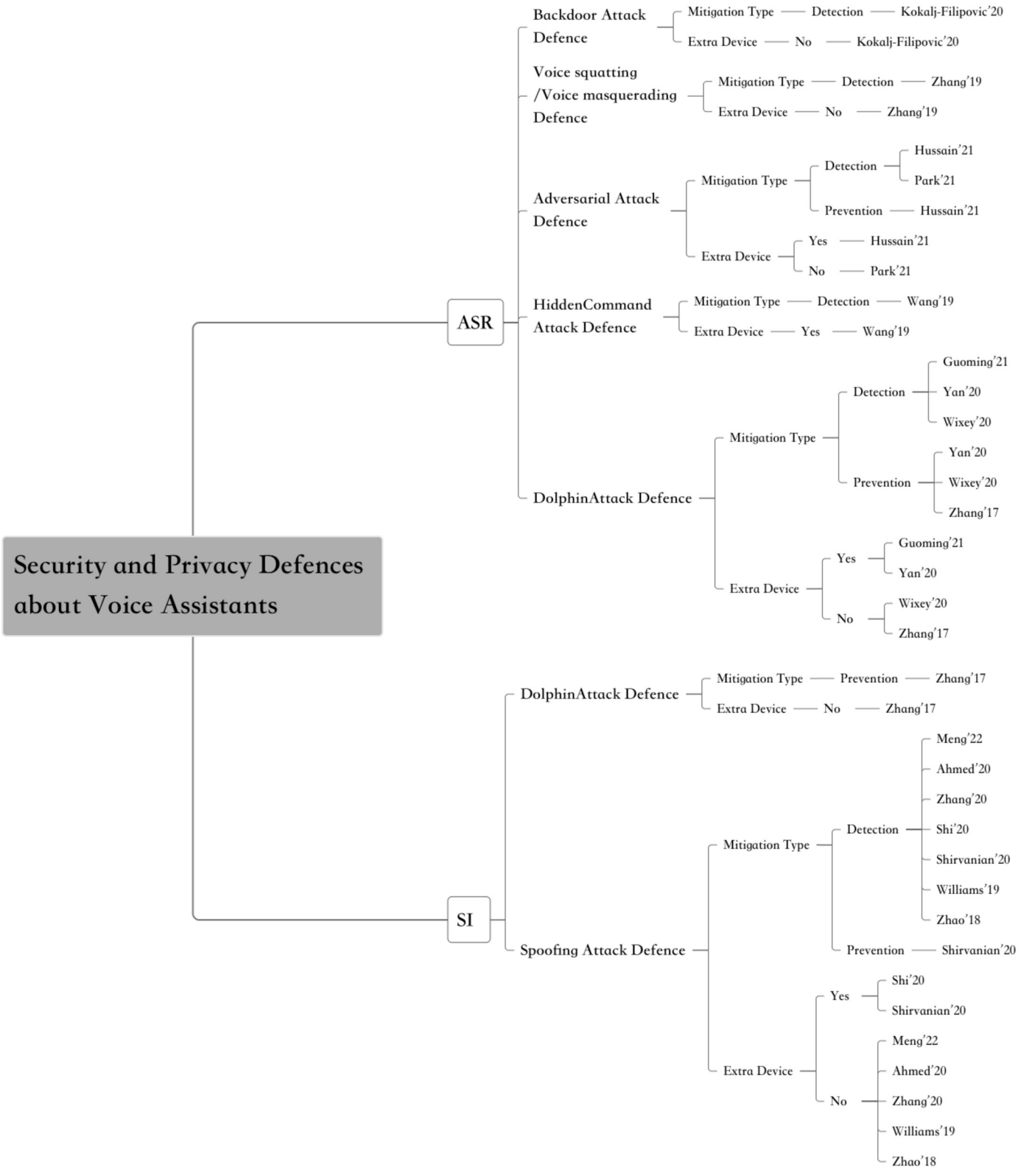

Fig. 5. Taxonomy of security and privacy defenses towards voice assistant applications.

Table 1
A comparison of existing attacks on ASR systems. The medium of a practical attack includes: A: Over-the-air, L: Over-the-line, P: Over-the-phone, S: Over-the-surface. The target model of a practical attack includes: I)Deep Speech, II)CNN based model, III) RNN based model, IV) DNN based model, V)Commercial smart devices. The setting of a practical attack includes: B: Black-box, W: White-box, G: Grey-box.

Type	Year	Paper	Medium	Goal	Target Model	Setting	Attack Approach	Success	Transferability
			A L P S		I II III IV V	B W G			
Backdoor	2021	Kasher'21	★ □ □ □	T	★ □ □ □ □	★ □ □	Comparison of Characteristics	50%	Y
		Koffas'21	★ □ □ □	T	□ ★ □ □ □	□ ★ ★	Dataset Poisoning	>=99%	PY
Voice squatting /Voice masquerading	2019	Zhang'19	★ □ □ □	T	□ □ □ □ ★	★ □ □	Use third party market	>=50%	-
Adversarial	2020	Serrano'20	□ ★ □ □	U	★ □ □ □ □	□ ★ □	Adversarial perturbations	-	PY
		Chen'20	★ □ □ □	T	★ □ □ □ □	□ ★ □	Domain adaptation algorithm	D, >=90%	-
	2019	Taori'19	□ ★ □ □	T	★ □ □ □ □	★ □ □	Combine genetic algorithms and gradient estimation	35%	-
	2018	Carlini'18	□ ★ □ □	T	★ □ □ □ □	□ ★ □	End-to-end	100%	PY
Hidden Command	2021	Abdullah'21	□ □ ★ □	U	★ □ □ □ ★	★ □ □	Interfere signal preprocessing	100%	Y
		Zhang'21	★ ★ □ □	T	□ □ □ □ ★	★ □ □	Faste speech	D, >=90%	Y
	2018	Schonherr'18	□ ★ □ □	T	□ □ □ ★ □	□ ★ □	Psychoacoustic hiding	98%	PY
	2019	Abdullah'19	★ ★ □ □	T	★ □ □ □ ★	★ □ □	Generic attack method	80%	Y
DolphinAttack	2020	Yan'20	□ □ □ ★	T	□ □ □ □ ★	★ □ □	Solid material	D, >=90%	-
		Wixey'20	□ ★ □ □	U	□ □ □ □ ★	★ □ □	Internet attack	-	PY
	2017	Zhang'17	★ □ □ □	T	□ □ □ □ ★	★ □ □	Modulate the frequency of the voice signal	D, >=90%	PY

★ Applicable □ Not applicable T: Targeted U: Untargeted D: Depending on the experimental environment Y: Yes PY: Presumably Yes - None (or unspecified)

Table 2
A comparison of existing defensive methods for ASR systems. The type of a defense includes: D: Detection, P: Prevention.

Mitigated Attacks	Year	Type	Paper	Methods	Extra Device	Success	Transferability
		D P					
Backdoor	2020	√ ×	Kokalj-Filipovic'20	Deep Learning Classifiers	×	-	Y
Voice squatting /Voice masquerading	2019	√ ×	Zhang'19	Context-sensitive detector	×	95%	-
Adversarial	2021	√ √	Hussain'21	LPC and Mel extraction inversion	√	-	Y
		√ ×	Park'21	Add noise to logits	×	-	PY
Hidden Command	2019	√ ×	Wang'19	Extract vibration features from motion sensor	√	100%	Y
DolphinAttack	2021	√ ×	Guoming'21	Decay rate difference	√	99%	Y
	2020	√ √	Yan'20	Monitor the characteristics of the source signal	√	-	-
		√ √	Wixey'20	Use imperceptible sounds as covert channels	×	-	Y
	2017	× √	Zhang'17	Hardware-based: microphone enhancement and baseband cancellation; Software-based: machine learning	×	-	PY

√ Positive × Negative - None (or unspecified) Y: Yes PY: Presumably Yes

privacy issues on a particular target model or offered new attack strategies. Table 1 and Table 3 provide additional details about each study that in the taxonomy Fig. 4, which can aid readers in understanding and comparing each work. We specifically provide the publication year, publication venue, learning task for the target model, attack knowledge available to the attacker, specific attack approach, baseline for the proposed attack, metrics for evaluating attack performance, and datasets used in the experiments for each paper listed in Tables 1 and 3. The correspondence defensive methods are summarized in Tables 2 and Table 4.

3. Attacks and defenses in ASR-based voice assistant applications

3.1. Backdoor attacks and defenses in ASR

The use of backdoor attacks against Automatic Speech Recognition (ASR) models is a common occurrence (Table 2). In these attacks, adversaries embed inaudible signals into data, such as music clips or voicemail messages, to disrupt the processing and decision-making of machine learning models. The objective of a backdoor attack is to insert a specific input, known as a trigger, into the model, resulting in the model making desired decisions based on that trigger. Adversarial audio is utilized to translate concealed orders into the attacker's intended command, exploiting the voice assistant's neural processing network while remaining imperceptible to human ears.

Kasher et al. Kasher et al. (2021) employed a backdoor system to control smart devices that respond to voice commands. They introduced potent, noise-resistant adversarial audio perturbations to music-only audio samples, effectively translating them into target commands. Their backdoor method exhibited efficacy when applied to both music and speech-based samples, enabling perturbations of various types. The study evaluated different base vectors, target words, and perturbation strengths to maximize the impact of the backdoor attack. Although the

selection of the target phrase is crucial, they achieved a transcription accuracy rate of over 50%.

Outsourcing the training process of ASR models or purchasing pre-trained models introduces vulnerabilities to various adversaries, including backdoor attacks. Koffas et al. Koffas et al. (2021) investigated backdoor attacks against ASR systems. They explored the injection of inaudible triggers, which make it challenging to detect backdoor attacks. The study utilized datasets with 10 and 30 classes, along with three neural networks, to assess the impact of trigger type, duration, and location. The findings revealed that launching an inaudible backdoor attack against ASR systems is relatively straightforward, requiring the attacker to poison only about 0.5% of the training samples. The duration of the trigger, even if inaudible, can significantly enhance the attack's success, and short non-consecutive triggers inserted into less than 0.5% of the training dataset can achieve attack success rates higher than 99%.

While the effectiveness of the attack strategies in these papers is evident, no defensive mechanisms were explored. Future research in the backdoor attack domain should focus on developing effective defenses. Kokalj-Filipovic et al. Kokalj-Filipovic et al. (2020) emphasized the need for further investigation into the detection of adversarial backdoor samples. They utilized raw acoustic data in covert attacks, employing inaudible messages transmitted through backdoor channels. By creating a backdoor dataset and analyzing the effects of the backdoor channel on acoustic data classification, they observed a decline in classifier accuracy. The extent of deterioration varied across deep classifier types and was less pronounced for classifiers trained with autoencoders. The authors suggested statistics, such as the empirical entropy layer of the classifier output or the log-likelihood of a variational autoencoder used for pre-training, that could be employed to identify out-of-distribution data generated by the backdoor channel. The initial findings underscore the necessity for further research before deploying deep learning classifiers as backdoor covert channel detectors.

In summary, backdoor attacks involve concealing inaudible signals within data to manipulate ASR models through covert backdoor channels, influencing the resulting transcriptions. These attacks are not limited to human voices and can be carried out using pure music clips or noisy audio that would not produce accurate transcriptions without perturbations. Moreover, the duration of the trigger signal can be as long as the injected audio, enhancing the attack's potency without revealing the signal. The recent backdoor attack techniques have achieved attack success rates exceeding 90% on average.

3.2. Voice squatting and voice masquerading

Voice assistants are a ubiquitous part of people's daily lives in the modern world. One must first authenticate to hear instructions and call other services using a voice assistant. As the third-party market grows, more and more service providers add new functions (skills) to it. As a result, attackers can trick voice assistants by uploading malicious skills to unofficial markets.

To evaluate how much impact the skills in the third-party market have on the security of voice applications, Zhang et al. Zhang et al. (2019) examine whether such remote, massive attacks through the third-party market are plausible. Voice squatting and masquerading were identified as two novel attack techniques. Voice squatting means imitating a legal ability by employing a wake phrase that sounds like the legal skill you intend to use. For instance, employing the lawful skill "open capital one" prevents voice assistants from utilizing the criminal skill "capital won". When you hear "capital one please," call another talent. When an attacker uses speech masquerading to communicate with a user and get sensitive information, they pose as a voice assistant or a called skill. The trials the authors did on the Google Home and Amazon Echo, which included user research and actual installations, revealed that they both offer a genuine threat.

Amazon and Google constructed a new squat detector available on Alexa and Google's marketplace. It confirms the importance of voice squatting and masquerading. A skill named scanner was created and available on the Amazon and Google Skills marketplaces to mitigate the risk. It identified several vulnerable Alexa skills and published skill names, indicating that the attack may have affected thousands of VPA customers. Additionally, they developed and used a context-sensitive detector to 95% accurately reduce the threat of voice masquerading. Future work will be needed to improve voice channel security and to verify the people engaged without interfering with the VPA system's availability.

3.3. Adversarial attacks and defenses in ASR

Recurrent neural networks (RNNs) are commonly utilized in time series machine learning systems, including Automatic Speech Recognition (ASR) systems. Serrano et al. Serrano et al. (2020) highlight the susceptibility of RNNs to periodic adversarial perturbations due to their unique memory and parameter-sharing capabilities. Their attack on the DeepSpeech model generates previously unknown cases in real-time by producing antagonistic perturbations. Experimental results demonstrate an average 0.426 cross-correlation and a mean square error of 0.017 between undisturbed and disturbed audio samples. The smoothed BLEU score between original and disturbed DeepSpeech transcripts is 0.221. While previous attacks on RNNs have focused on white-box and black-box iterative approaches, this strategy exploits the temporal nature of the problem, leveraging RNNs' real-time adversarial perturbation generation capabilities.

Carlini et al. Carlini and Wagner (2018) conduct an end-to-end adversarial attack against the DeepSpeech model in a white-box setting. Despite the difficulty of optimizing with Mel-frequency spectral coefficient (MFCC) preprocessing, they directly feed unprocessed data as input to the classifier. Their attack is 100% effective regardless of the required transcription or source audio samples. By embedding speech within audio that should not be identified as speech, they demonstrate the potential to transcribe speech into music, conceal speech from transcription, and achieve a transcription rate of up to 50 characters per second. This highlights the value of targeted audio adversarial samples in automated speech recognition, revealing that linearity does not hold in the audio domain.

Taori et al. Taori et al. (2019) focus on improving adversarial attacks on deep recurrent networks in ASR systems, particularly in black-box environments. They employ a combination of evolutionary algorithms and gradient estimation techniques to generate adversarial samples. After 3,000 generations, they achieve a 35% target attack success rate, maintaining 94.6% audio file similarity and 89.25% target attack similarity. The findings suggest that combining genetic algorithms with gradient estimation produces superior adversarial samples compared to using either approach alone. Their method successfully targets deep nonlinear ASR systems, generating near-perfect transcriptions while maintaining a high degree of similarity.

Chen et al. Chen et al. (2020) address the challenge of signal distortion during aerial transmission by proposing the generation of inaudible sounds that can withstand air transmission. They leverage the frequency selectivity of devices and channels, employing a two-phase architecture called Metamorph. This approach creates an initial perturbation based on prior measurements that account for core distortion effects. A domain adaptation technique is then applied to enhance the attack range and reliability. Evaluation results demonstrate a high attack success rate of 90% at close attack ranges of 6 meters.

In summary, deep learning-based ASR systems face significant security and privacy challenges, particularly in relation to adversarial attacks. These attacks have demonstrated success rates exceeding 90% on average. Researchers have explored various attack strategies, including exploiting RNN vulnerabilities, end-to-end adversarial attacks, targeted attacks in black-box environments, and addressing signal dis-

tortion during aerial transmission. The development of robust defense mechanisms against adversarial attacks remains an important area for future research.

3.4. Hidden command attacks and defenses in ASR

Deep neural networks (DNNs) have significantly advanced Automatic Speech Recognition (ASR) by approximating human hearing and understanding. However, DNNs are highly susceptible to adversarial perturbations. One type of attack is the hidden command attack, which exploits imperceptible perturbations that are correctly understood by ASR but not by human listeners.

Schonherr et al. Schönherr et al. (2018) introduced an innovative adversarial example based on psychoacoustic concealment. Their attack utilized DNN-based ASR systems, incorporating a back-propagation stage into the initial analytic process. The attack achieved a success rate of 98% in less than two minutes for 10-second audio recordings. Human listeners were unable to comprehend the intended transcription, although the accuracy remained unchanged.

The DNN-HMM system underwent forced alignment and backpropagation, generating undetectable adversarial perturbations with high reliability. The study by Schonherr et al. Schönherr et al. (2018) explored algorithm variables such as the number of iterations and permitted hearing threshold deviations. Their approach resulted in less distortion compared to previous studies, enabling targeted adversarial situations. Further research should incorporate psychoacoustic models to strengthen ASR systems against these simple attacks. Evaluating attacks in real-world scenarios and commercial ASR systems using black-box settings is also essential.

Abdullah et al. Abdullah et al. (2019) addressed the limitations of white-box attacks by utilizing a model-agnostic (black-box) approach in hidden command attacks. They leveraged signal processing methods commonly used by voice processing systems (VPS) to generate data for input into machine learning systems, making hidden command attacks more feasible. Their study evaluated 12 machine learning models, including proprietary ones, against various targets. They demonstrated the effectiveness of four classes of perturbations resulting in unintelligible sounds. The attacks performed well across different hardware setups, emphasizing the role of domain-specific knowledge in successful covert voice command attacks.

Building upon this research, Abdullah et al. Abdullah et al. (2021) explored pipeline phase attacks as black-box and transferrable attacks. By modifying a few audio frames, the attack achieved a 100% mistranscription and misrecognition rate. These attacks exploited the model's sensitivity to minor, imperceptible speech components critical for accuracy. Certain English phonemes, especially vowels, were found to be more vulnerable. These attacks proved effective in cellular networks, where transcoding, jitter, and packet loss weaken the signal. Unperceivable phrases or signals with a significant impact on ASR or speaker identification (SI) systems can be utilized to construct more potent models for reducing ASR and AVI system vulnerabilities. Adversarial training can provide partial mitigation, but stronger defenses are ultimately necessary to counter these attacks.

Zhang et al. Zhang et al. (2021) proposed a novel attack independent of the target model, exploiting rapid speech that often leads to misinterpretation by both humans and ASR systems. They manipulated the phonetic structure of target voice commands in high-speed versions to deceive ASR systems into inferring secret meanings. The attack consistently succeeded across seven real-world ASR systems in various settings, achieving a high success rate for adversarial commands. This study highlighted the vulnerability of current ASR systems to fast speech and demonstrated the efficacy of the proposed CommanderGabble method.

Defending against hidden command attacks poses challenges, resulting in fewer research outcomes on defenses compared to attacks. Wang et al. Wang et al. (2019) proposed a detection technique that identified and analyzed distinctive audio signatures of spoken instructions transmitted through vibration. They exploited audio-induced surface vibrations detected by motion sensors, which are difficult to replicate, to conceal verbal orders and trick authentication systems. Their learning-based method distinguished conventional voice instructions from covert ones using temporal and frequency-domain statistical characteristics and acoustic signals extracted from motion sensor data. Experimental results showed 99.9% accuracy in differentiating disguised from conventional voice instructions using low-cost motion sensors with low sample frequency. The system also benefited from speaker motion sensors.

Nevertheless, Wang et al. Wang et al. (2019) acknowledged potential issues with the system's playback procedure, such as front-end mode incursion and playback delays. Modulating the front-end playback sound to inaudible frequencies could potentially achieve zero incursion. Combining front-end and rear-end playback configurations using hidden and inexpensive devices with built-in speakers and motion sensors has been explored. The unsupervised learning-based strategy could be extended to protect against additional threats, such as ultrasonic attacks, with minimal training. Further investigation into separating playback noises and human voices in the vibration domain would also be worthwhile.

In conclusion, hidden command attacks exploit imperceptible perturbations that fool ASR systems but remain undetected by humans. Various studies have explored different attack methods, including psychoacoustic concealment, model-agnostic approaches, and exploiting fast speech. Defending against these attacks requires robust defenses, and research should focus on enhancing ASR systems' resilience while considering real-world scenarios and commercial systems. Additionally, methods utilizing distinctive audio signatures and vibration analysis show promise in detecting hidden command attacks, although further refinements and investigations are needed.

3.5. DolphinAttack and its defenses in ASR

Speech assistants are susceptible to transcribed signal injection at inaudible frequencies. In earlier studies, high-frequency signal injection into audio has received much attention. Sound waves above or below the human hearing range are used as an attack strategy in DolphinAttack. There are many ways to attack voice assistants, and research has concentrated on inserting secret orders into audio samples. Zhang et al. Zhang et al. (2017) are the first to propose a silent attack named DolphinAttack. Although DolphinAttack uses an ultrasonic carrier wave (with a frequency over 20 kHz) to avoid human perception, voice assistants likely discern the signal because of the hardware properties of the microphone. DolphinAttack was successfully conducted on several voice recognition systems, including Siri, Google Now, Samsung S Voice, Huawei HiVoice, Cortana, and Alexa.

Zhang et al. Zhang et al. (2017) suggested baseband cancellation and microphone augmentation as two hardware-based defense tactics. In terms of microphone improvement, an improved microphone suppresses acoustic sounds with ultrasonic-range frequencies. Signals in the ultrasonic frequency range with AM modulation properties can be identified and demodulated to obtain baseband in terms of baseband cancellation. Since there will be no association between the acquired audible speech signal and the noise in the ultrasonic region, a command cancellation procedure will not affect the microphone's working condition.

However, it is challenging to stop DolphinAttack. To launch DolphinAttack, attackers surreptitiously insert malicious orders into voice assistants and control systems by modulating audible speech with ultrasonic waves (such as doors or smart speakers). Guoming et al. Guoming et al. (2021) proposed EarArray as a simple technique for detecting DolphinAttack. EarArray exploits the fact that ultrasonic waves decay more quickly than audible sound to identify DolphinAttack. EarArray compares the command sound signal with the decay rate using nu-

Table 3
A comparison of existing attacks on SI systems. The medium of a practical attack includes: A: Over-the-air, L: Over-the-line, P: Over-the-phone, S: Over-the-surface. The goal of a practical attack includes: T: Targeted, U: Untargeted. The target model of a practical attack includes: I)Deep Speech, II)CNN based model, III) RNN based model, IV) DNN based model, V)Commercial smart devices. The setting of a practical attack includes: B: Black-box, W: White-box, G: Grey-box.

Type	Year	Paper	Medium	Goal	Target Model	Setting	Attack Approach	Success	Transferability
			A L P S	T U	I II III IV V	B W G			
Adversarial	2021	Chen'21	★ ★ □ □	★ ★	□ □ □ □ ★	★ □ □	Optimization	99%	PY
Backdoor	2020	Zhai'20	□ ★ □ □	★ □	□ □ □ ★ □	★ □ □	Clustering-based attack	45%	Y
hidden command	2021	Abdullah'21	□ □ ★ □	□ ★	★ □ □ □ ★	★ □ □	Interfere signal preprocessing	100%	Y
hidden command	2019	Abdullah'19	★ ★ □ □	★ □	□ □ □ □ ★	★ □ □	Generic attack method	80%	Y
DolphinAttack	2017	Zhang'17	★ □ □ □	★ □	□ □ □ □ ★	★ □ □	Modulate the frequency of the voice signal	D	PY

★ Applicable □ Not applicable D: Depending on the experimental environment Y: Yes PY: Presumably yes

merous microphones integrated into the smart device. EarArray could determine the attacker's direction with 97.89% accuracy and detect inaudible spoken orders with 99.0% accuracy.

The viability of transmitting silent ultrasonic attacks using solid materials is proposed as SurfingAttack in Yan et al. Yan et al. (2020). Unlike wireless transmission in previous studies, the new attack may conceal itself inside or beneath the solid material, opening up a new path for inaudible attacks. SurfingAttack employs the energy transmission mechanism of ultrasonic guided waves, which is a viable and affordable attack method. SurfingAttack effectively attacked devices from 30 feet away while using just 0.75 W of attack signal power. Several trials were tested to determine the scope and boundaries of this hazard. The voice response was listened to at a low volume to facilitate dialogue between an opponent and the voice-controllable device. SurfingAttack may allow attackers to decrypt SMS passwords or place phony calls.

Yan et al. Yan et al. (2020) proposed several strategies for defense against SurfingAttack. There are three recommendations to reduce or eliminate acoustic vibrations in the ultrasonic range, including improving the microphone's hardware arrangement, positioning the gadget atop a soft woven fabric, and differentiating the frequency difference between attacks and normal signals. 54 attacks were manifested using various attack settings (i.e., frequency, table material, distance, baseband signal, and device). Because the human voice contains few extremely high-frequency components, the received voice signal will be labeled an attack if the attack index is higher than the pre-set threshold. This defense technique may fail if the device's audio low-pass filter has a cut-off frequency lower than 10 kHz because of inadequate data for attack index calculation.

The network attacks in Wixey et al. Wixey et al. (2020) cause smart devices to create high-frequency (17-21 kHz) independently and low-frequency (60-100 Hz) sounds, transforming them into acoustic attack weapons. Several gadgets appeared capable of emitting noises at volumes that matched or surpassed several advised limits. The measurements were conducted in an anechoic chamber at a distance of one meter. It can impact individuals across a wide region and be used for large and lethal devices. For instance, a speaker system in a car or a linked PA system during a concert or sporting event may be attacked to emit dangerous noise to human beings. Other, "noisier" channels may also be used, including smart TV broadcasts and injecting HFN or LFN into phone conversations.

Wixey et al. Wixey et al. (2020) evaluated two free Android applications with an external microphone to generate an alarm when the sound intensity increased, particularly HFN. Other defense tactics against Wixey's attack include routing inaudible noises, restricting speakers' frequency range, notifying the user, disabling playing audio files outside the audible range, and strengthening mobile app permissions. Consumer and business antivirus detection engines can incorporate heuristics to find these threats. A confirmation prompt may be sent to the user to ask if they want to proceed if, for instance, a combination of specific behaviors is frequently identified by antivirus engines as suspicious behavior.

4. Attacks and defenses in SI-based voice assistant applications

Most products combine ASR and SI models in the voice assistant applications market. Some applications only provide speech recognition functions. However, stand-alone SI function applications are rare. The research in this area has fewer works that target SI systems only. The number of attack kinds used on SI systems is also fewer than in ASR systems. This section introduces attack and defense techniques for SI systems.

Tables 3 and 4 present notable investigations of attacks against Speaker Identification (SI) systems and defenses employed for SI systems. The attack studies are compared based on the target model, attack methods, and attack settings. Similarly, the defense studies are compared according to their defense methods, the need for additional devices, success rates, and related factors. In order to obtain outcomes that are both realistic and applicable, all experiments were conducted in varying levels of ambient noise, replicating real-life environments. The majority of these studies utilized the widely accessible open-sourced TIMIT dataset, which is a prominent collection of speech data. A few studies, however, opted to gather their own datasets through participant contributions.

4.1. Attacks against SI-based models

The vulnerability of Speaker Identification (SI) systems to adversarial attacks is a significant concern in IoT devices, where SI systems are commonly employed for biometric identification and authentication. While research has made progress in understanding adversarial attacks in white-box settings, the challenge of adversarial attacks in the black-box context remains open.

In 2021, Chen et al. Chen et al. (2021a) conducted an extensive investigation on adversarial attacks against SI systems in black-box environments, making notable contributions to this field. They introduced a novel adversarial technique called FAKEBOB, which generates adversarial samples by optimizing adversarial sound intensity and imperceptibility. The authors proposed a technique to estimate the score threshold component of SI systems and used it to address optimization challenges. FAKEBOB achieved a high success rate of 98% in 16 attack scenarios and demonstrated the difficulty for human listeners to distinguish between neutral and aggressive speakers. Furthermore, FAKEBOB rendered several countermeasures against adversarial attacks in speech recognition ineffective.

The reliance on training data collected from third parties exposes SI systems to security risks. Backdoor attacks, where an attacker poisons the training data to introduce a secret backdoor into the speaker verification model, pose a significant threat. Zhai et al. Zhai et al. (2021) proposed a cluster-based attack strategy, using various triggers for poisoned samples from different clusters. This approach successfully evaded prior backdoor defenses by employing unregistered defined triggers for model validation. Experimental evaluations across datasets showed Attack Success Rates (ASR) greater than or equal to 45% for

Table 4
A comparison of existing defensive methods for SI systems. The type of a defense includes: D: Detection, P: Prevention.

Mitigated Attacks	Year	Type		Paper	Methods	Extra Device	Success	Transferability
		P	D					
Spoofing	2022	×	√	Meng'22	Array Fingerprint	×	99.84%	-
	2020	×	√	Ahmed'20	Compare the spectral power difference between real human speech and the replayed speech played through speakers	×	91.30%	PY
		×	√	Zhang'20	Catch the dissimilarities between bone-conducted vibrations and air-conducted voices when human speaks	×	97%	N
		×	√	Shi'20	Cross-domain comparisons: Audio&vibration	√	97.20%	Y
		√	√	Shirvanian'20	The WER for the synthesized voices are 2-3 times more than the WER for natural voices	√	>=95%	-
	2019	×	√	Williams'19	Combine x-vector attack embeddings with signal processing features	×	-	-
	2018	×	√	Zhao'18	Weighting framework with a Gaussian Mixture Model (GMM) classifier	×	>=98%	PY
DolphinAttack	2017	√	×	Zhang'17	Hardware-based: microphone enhancement and baseband cancellation; Software-based: machine learning	×	-	PY

√ Positive × Negative - None (or unspecified) Y: Yes PY: Presumably Yes

the tested speaker verification techniques. Detecting this attack is challenging, as the tainted data's Equal Error Rate (EER) is similar to that of a model trained with a clean dataset. In real-world applications with multiple users, the resulting ASR is further amplified, highlighting the need for robust verification techniques.

These findings shed light on the vulnerability of SI systems to adversarial attacks, particularly in black-box scenarios. The FAKEBOB technique demonstrates the effectiveness of adversarial attacks in evading countermeasures, while the cluster-based backdoor attack strategy provides a fresh perspective for creating new attacks. Addressing these challenges is crucial for enhancing the robustness and security of speaker verification techniques.

4.2. Defensive methods for SI-based models

Speech assistant systems are susceptible to acoustic attacks, including spoofing attacks, because voice signals are accepted in open spaces and channels. Modern VUIs that employ classic voice authentication techniques are susceptible to spoofing attacks, in which an evil adversary impersonates a real user by speaking commands that have already been recorded or synthesized. Once the voice assistant's SI system has been tricked, some risky actions, like making bulk purchases and phoning friends and family, may be executed.

Many studies have focused on addressing the vulnerability of Speaker Identification (SI) systems to spoofing attacks. Notably, the use of Constant-Q spectral coefficients (CQCC) and scattered spectral coefficients (SCC) has been effective in speech synthesis (SS) and speech conversion (VC) to identify fake speech signals. However, the equal error rate (EER) remains high for certain types of attacks, leading to selective detection degradation. To mitigate this, Zhao et al. Zhao et al. (2018) proposed an independent detector with adaptive weighting. Their approach combines CQCC and SCC features at the score level, and employs a new clustering technique to assess data structure. By selecting appropriate weighting variables based on the clustering characteristics of truthful and dishonest subgroups, they achieved lower EERs using a Gaussian Mixture Model (GMM) classifier on the ASVspoof 2015 database.

Williams et al. Williams and Rownicka (2019) introduced a novel detection mechanism that combines x-vector attack embeddings with signal processing characteristics. Their system employs convolutional neural networks (CNNs) and spoken audio representations, generating x-vectors using Mel Frequency Cepstral Coefficients (MFCCs) through a Time Delay Neural Network (TDNN). The study includes diverse attack types and contexts, and augments the data with Gaussian noise to improve resistance against unseen attacks. The system's performance is evaluated using the Tandem Detection Cost Function (tDCF) and Equal Error Rate (EER). The use of frame-level 40-dim MFCC features without frequency range restriction could be further explored to capture variations in acoustic settings.

Shirvanian et al. Shirvanian et al. (2020) proposed a mitigation strategy for increasing the security of speaker verification from voice synthesis attacks without additional hardware. They leveraged speech transcribers, as synthetic speech is typically transcribed less accurately compared to genuine voices. By discarding terms not found in the reference dictionary and accepting transcribed text with a specific number of word errors, their detection technology achieved low false rejection rates and false accept rates for phonetically unique terms. However, this strategy may be ineffective against sophisticated speech synthesis technology that can achieve transcription accuracy similar to natural audio.

Shi et al. Shi et al. (2020) presented a training-free voice authentication system called WearID. It utilizes cross-domain voice similarity between audio recorded by wearable accelerometers and voice assistant (VA) systems' microphones. This approach provides increased security for speaker verification in VA systems without requiring active user participation or the storage of privacy-sensitive voice samples. WearID utilizes a special vibration-sensing interface and transforms microphone data into low-frequency "motion sensor data" for domain comparisons. The system demonstrates high accuracy in identifying users' voice commands in regular use and detecting fraudulent voice commands in audible/inaudible attacks.

Zhang et al. Zhang et al. (2020) developed a continuous liveness detection technique for secure Voice User Interfaces (VUIs) in IoT contexts. It distinguishes between air-conducted vocals and bone-conducted vibrations during speech, validating real users and identifying spoofing attempts. The technique operates without additional software or hardware beyond standard loudspeakers and microphones. By utilizing the weaknesses of VUIs as a detection method and probing with ultrasound, this technique can identify spoofing attempts using replay attacks.

Ahmed et al. Ahmed et al. (2020) proposed a voice live detection system called "Void" to mitigate replay attacks. Void compares spectral power differentials between actual human speech and replayed speech played over speakers to identify spoofing attacks. Their system achieves low error rates using a single classification model with 97 features, consuming significantly less memory and providing faster detection compared to conventional approaches. By combining Void with a Gaussian mixture model employing Mel Frequency Cepstral Coefficients (MFCCs), the error rate can be further reduced.

Meng et al. Meng et al. (2022) developed an active feature fingerprint utilizing the microphone array in smart speakers for source identification. They demonstrated the robustness of the array fingerprint to environmental changes and human mobility, leveraging the circular architecture of the microphones. The proposed ARRAYID approach achieves high accuracy in passive liveness detection, surpassing other techniques, by incorporating characteristics that cooperate with array fingerprints.

In summary, while certain studies focused on specific papers without delivering clear insights, these aforementioned works address the vulnerability of SI systems to spoofing attacks and propose mitigation strategies through various techniques such as adaptive weighting, x-vector embeddings, speech transcribers, cross-domain similarity, vibration-based detection, spectral power differentials, and active feature fingerprints.

5. Privacy issues beyond technical attacks

Users of IoT devices may not always be aware of who can access their recordings, as many smart products, including TVs, doorbells, and more, have built-in microphones. Protective Jamming Devices (PJDs) are commonly used to prevent eavesdropping on conversations by placing them on top of smart voice assistants' speakers. However, even with PJDs, there is still a risk of voice eavesdropping due to advanced signal processing techniques used by modern voice assistants that reduce background noise and enhance speech. Hackers could potentially disrupt user speech by gaining access to recordings made by smart speakers.

Walker et al. Walker and Saxena (2021) investigate the effectiveness of protective jamming devices in preventing eavesdropping on conversations. They analyze the impact of white Gaussian noise as a PJD interferer and evaluate its effectiveness using extensive experiments and signal processing techniques.

Mitev et al. Mitev et al. (2020) propose LeakyPick, an architecture that monitors the communication flow of smart devices in the cloud to prevent private user conversations from being recorded. They analyze encrypted MAC-layer communication and identify potential vulnerabilities in voice assistants' wake word detection.

Young et al. Young et al. (2022) focus on protecting young users from harmful third-party features in voice applications. They conduct a dynamic analysis to identify policy-violating voice apps, providing insights for platform providers to prevent the release of such applications. Le et al. Le et al. (2022) develop a natural language processing system to analyze voice assistant dialogues and identify risky content for children. They highlight the potential risks associated with voice applications targeting kids and the need for stricter regulations and examinations.

He et al. He et al. (2022) investigate the distinctiveness of voiceprints based on different words and explore their applications in wake-word selection and enhancing voice assistant safety. Ahmed et al. Ahmed et al. (2022) propose EKOS as a remedy to protect against unintentional and adversarial activations during the wake-up procedure of voice assistants. EKOS leverages spatial redundancy in the auditory environment to reduce unexpected activations while maintaining accuracy. Chen et al. Chen et al. (2021b) address the FakeWake phenomena, focusing on the inadvertent triggering of voice assistants by fuzzy words. They propose a systematic framework for generating and understanding fuzzy words, analyzing the causes of false acceptance by wake-up detectors, and proposing mitigation techniques.

There is also some work addressing some application logic issues. Cheng et al. Cheng et al. (2019) investigate incorporating additional data into acoustic signals for voice assistant processing, introducing a tagging technique to enhance privacy when using voice assistants.

Liao et al. Liao et al. (2020) examine the effectiveness of privacy policies in voice applications, revealing discrepancies and inaccuracies in existing regulations and highlighting the need for improved privacy protection measures.

In conclusion, the discussed research addresses various concerns related to voice assistants, including eavesdropping risks, backdoor attacks, privacy policies, harmful content in voice applications, voiceprint distinctiveness, and unintentional and adversarial activations. These studies provide valuable insights into protecting user privacy, improving security measures, and enhancing the overall safety of voice assistant technology.

6. Challenges and future research directions

Because the backdoor attack's performance has been relatively good, future studies should concentrate more on defensive strategies. Identifying the backdoor channels has always been a priority when mitigating backdoor attacks. Most existing defenses against backdoor attacks rely on identifying tainted inputs, which is not always reliable because of well-hidden triggers in sparse data. Retraining is resistant to backdoor attacks, but it is not practical because of its high computational time and indefinite calls to clean training data. In addition, it is presumable that the attacker taints the data in a "normal" manner by altering the training dataset. If the dataset is poisoned from a distance, it remains unknown whether such an inaudible attack is effective.

Several facets of voice squatting and voice masquerade attacks may benefit future research. These two strikes are new and constantly evolving due to the quickly expanding third-party skill market. The main obstacle to these attacks is the development of an autonomous retrieval system to search for prospective "susceptible" talents to hijack or fake. The existing approach performs admirably, but the searching step places undue reliance on human effort.

Compared to other attacks, the adversarial attack is one of the most developed attacks in speech. For the adversarial attack, there are still restrictions and difficulties. First, in contrast to the Dolphine Attack, the disturbance introduced would change the audio file and be audible to human hearing. In other words, a human defense might identify the antagonistic samples if he/she listened to every audio clip played. Thus, improving the transcript's plausibility without compromising its correctness remains a future work. It would stop human defenders from being suspicious enough to inspect the audio file. Second, the majority of adversarial attack methods include teaching the model to discover how to swap out or substitute certain target phrases, but it typically fails when encountering untrained or unseen samples. Third, unlike white box settings, the black box setting in adversarial attacks is significantly less investigated.

Most research in adversarial attacks focused on CNN and DNN models. For other machine learning areas, combining transformer ideas into their existing models to improve efficiency is a mainstream trend for end-to-end models. In the last four years, there has been little progress using the advantage of transformer models, except Carlini et al. Carlini and Wagner (2018) in 2018. Even with an impressive attack success rate, Carlini's study was based on white-box knowledge, which did not hugely raise the attack's performance compared to other white-box research. The defense side of the adversarial attack has been greatly developed over the years. Future studies should consider if original sentences can be extracted from adversarial samples after adding noise to lower the adversarial samples' accuracy, for example, Park et al. Park et al. (2021) in 2021. After testing various Gaussian noise distributions, they could retrieve the original transcription findings when the perturbation signal in the opposing samples was minimal. The sound is louder while making huge perturbations of adversarial samples to retrieve the actual transcription from the logit noise. Additional research is required to reconstruct the original transcript using logit noise.

Regarding hidden command attacks, the method proposed in Schönherr et al. (2018) has achieved good attack rates using psychoacoustic concealment. A good attack rate was maintained without affecting human understanding of the original speech samples. However, its main problem is that the injecting sample took a relatively long time compared to other methods. Moreover, it was not tested on black-box ex-

perimental environments. The transferability and robustness need to be enhanced. To make the attack more stable and transferable, future research should focus on decreasing the influence of adversarial training mitigation and avoiding transition loss over distance and medium. Another problem in this area is that the defense side was much weaker than the attack, making attacks more dangerous. With the very effective defense method introduced, voice attackers would face a challenge since it calls for them to imitate acoustic fingerprints in both the audio and vibration domains. The playback procedure needed in the detection process might result in front-end mode incursion and some playback delays. Future research may investigate whether and how playback noises and human voices are separated in the vibration domain.

A dolphin attack takes advantage of the microphone's hardware flaw, making it difficult to mitigate without hardware modification. Future research should pay attention to how to change the structure of the microphone and make the change more adaptable and cheaper to adapt. Dolphin attack was introduced in 2017 Zhang et al. (2017). For mitigation, some researchers have raised some ideas of modifying the existing microphone, for example, baseband cancellation and microphone augmentation. However, none of the defenses was adopted by the market. One reason could be that the defense only applies to single-attack mitigation. It is not worthwhile if the dolphin attack was used on a small scale. On the attack side, the attack over the air was mature enough; thus, much research focused on different transmission mediums. With good performance on solid medium, this attack was too specific and could only target specific users, and the attacker himself must have physical access to the attack environment or be close enough to the targeted object.

One of the most significant problems for the third-party market is the policy restriction to protect young users. With increasingly more young users storming into the voice assistant market and the average age of the users being increasingly younger, there are few regulations to distinct the accessibility of the content for adults and kids. The legibility of the skills released to the third-party market was not thoroughly checked due to a lack of restrictions for the service providers. The enforcement of automatic checks for the legibility of every skill before publishing is necessary to safeguard voice assistant applications. Adding a kid's friendly mode and requiring parents' authorization when detecting young users is also essential for voice assistant applications.

In summary, a few aspects should be refined to mitigate technical and non-technical threats in voice assistant applications. This section reviewed challenges and possible future research directions in backdoor attacks, voice squatting and voice masquerade attack, adversarial attack, hidden command attack, and dolphin attacks. The challenges and open issues include attack and defense sides for each attack. This section reviewed the security and privacy problem beyond technical attacks. Policy and supervision of the market are also important for protecting users.

7. Conclusion

An in-depth analysis of voice assistant security is provided in this article, with an emphasis on the attacks that may cause a voice assistant to act maliciously and the defenses against such attacks. We start by outlining the overall organization and process of a voice assistant. Then, based on the attacker's intention, threat model, attack strategy, and actual effectiveness, we briefly review several attacks. The six different technological attacks — backdoor attack, spoofing attack, adversarial attack, hidden command attack, and dolphin attack—are systematized and expanded upon. We categorize current countermeasures into two groups — detection and prevention—and organize them according to the types of assaults they were intended to avoid and their usefulness in terms of implementation costs, usability, and effectiveness. Then We outlined the wake-up word issue and application logic issues within voice assistants. We also discuss the prospective avenues and current difficulties in future research.

CRediT authorship contribution statement

Jingjin Li: Conceptualization(equal); Writing - Original Draft Preparation(lead); Writing - Review & Editing(lead). Chao Chen: Conceptualization(lead); Project Administration(lead); Resources(lead); Supervision(lead); Review & Editing(equal). Mostafa Rahimi Azghadi: Supervision(equal); Review & Editing(equal). Hossein Ghodosi: Supervision(equal); Review & Editing(equal). Lei Pan: Writing - Review & Editing(equal). Jun Zhang: Supervision(equal); Review & Editing(equal).

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

Data will be made available on request.

References

Abdullah, H., Garcia, W., Peeters, C., Traynor, P., Butler, K.R., Wilson, J., 2019. Practical hidden voice attacks against speech and speaker recognition systems. preprint. arXiv:1904.05734.

Abdullah, H., Rahman, M., Garcia, W., Warren, K., Yadav, A., Shrimpton, T., Traynor, P., 2021. Hear "no evil" see "kenansville". In: Efficient and Transferable Black-Box Attacks on Speech Recognition and Voice Identification Systems, pp. 712–729.

Ahmed, M.E., Kwak, I.Y., Huh, J.H., Kim, I., Oh, T., Kim, H., 2020. Void: a fast and light voice liveness detection system. In: Proceedings of the 29th USENIX Security Symposium (USENIX Security 20). USENIX Association, pp. 2685–2702. https://www.usenix.org/conference/usenixsecurity20/presentation/ahmed-muhammad.

Ahmed, S., Shumailov, I., Papernot, N., Fawaz, K., 2022. Towards more robust keyword spotting for voice assistants. In: Proceedings of the 31st USENIX Security Symposium (USENIX Security 22). USENIX Association, Boston, MA. https://www.usenix.org/conference/usenixsecurity22/presentation/ahmed.

Carlini, N., Wagner, D., 2018. Audio adversarial examples: targeted attacks on speech-to-text. In: Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW). IEEE, pp. 1–7.

Chen, G., Chen, S., Fan, L., Du, X., Zhao, Z., Song, F., Liu, Y., 2021a. Who is real bob? Adversarial attacks on speaker recognition systems. https://doi.org/10.1109/SP40001.2021.00004.

Chen, T., Shangguan, L., Li, Z., Jamieson, K., 2020. Metamorph: injecting inaudible commands into over-the-air voice controlled systems. In: Proceedings of the NDSS.

Chen, Y., Bai, Y., Mitev, R., Wang, K., Sadeghi, A.R., Xu, W., 2021b. Fakewake: understanding and mitigating fake wake-up words of voice assistants. In: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. Association for Computing Machinery, New York, NY, USA, pp. 1861–1883.

Cheng, P., Bagci, I.E., Yan, J., Roedig, U., 2019. Smart speaker privacy control - acoustic tagging for personal voice assistants. In: Proceedings of the 2019 IEEE Security and Privacy Workshops (SPW), pp. 144–149.

Cheng, P., Roedig, U., 2022. Personal voice assistant security and privacy—a survey. Proc. IEEE 110, 476–507. https://doi.org/10.1109/JPROC.2022.3153167.

Guoming, Z., Ji, X., Li, X., Qu, G., Xu, W., 2021. Eararray: defending against dolphinattack via acoustic attenuation. https://doi.org/10.14722/ndss.2021.24551.

He, R., Ji, X., Li, X., Cheng, Y., Xu, W., 2022. "OK, siri" or "hey, google": evaluating voiceprint distinctiveness via content-based PROLE score. In: Proceedings of the 31st USENIX Security Symposium (USENIX Security 22). USENIX Association, Boston, MA. https://www.usenix.org/conference/usenixsecurity22/presentation/he-ruiwen.

Kasher, M., Zhao, M., Greenberg, A., Gulati, D., Kokalj-Filipovic, S., Spasojevic, P., 2021. Inaudible Manipulation of Voice-Enabled Devices Through BackDoor Using Robust Adversarial Audio Attacks: Invited Paper. Association for Computing Machinery, New York, NY, USA, pp. 37–42.

Koffas, S., Xu, J., Conti, M., Picek, S., 2021. Can you hear it? Backdoor attacks via ultrasonic triggers. preprint. arXiv:2107.14569.

Kokalj-Filipovic, S., Kasher, M., Zhao, M., Spasojevic, P., 2020. Detecting acoustic backdoor transmission of inaudible messages using deep learning. In: Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning. Association for Computing Machinery, New York, NY, USA, pp. 80–85.

Le, T., Huang, D.Y., Apthorpe, N., Tian, Y., 2022. Skillbot: identifying risky content for children in alexa skills. ACM Trans. Internet Technol. 22, 1–31.

Liao, S., Wilson, C., Cheng, L., Hu, H., Deng, H., 2020. Measuring the effectiveness of privacy policies for voice assistant applications. In: Proceedings of the Annual Computer Security Applications Conference. Association for Computing Machinery, New York, NY, USA, pp. 856–869.

Meng, Y., Li, J., Pillari, M., Deopujari, A., Brennan, L., Shamsie, H., Zhu, H., Tian, Y., 2022. Your microphone array retains your identity: a robust voice liveness detection system for smart speakers. In: Proceedings of the 31st USENIX Security Symposium (USENIX Security 22). USENIX Association, Boston, MA. https://www.usenix.org/conference/usenixsecurity22/presentation/meng.

Mitev, R., Pazii, A., Miettinen, M., Enck, W., Sadeghi, A.R., 2020. Leakypick: Iot audio spy detector. In: Proceedings of the Annual Computer Security Applications Conference. Association for Computing Machinery, New York, NY, USA, pp. 694–705.

Park, N., Ji, S., Kim, J., 2021. Detecting audio adversarial examples with logit noising. In: Proceedings of the Annual Computer Security Applications Conference, pp. 586–595.

Schönherr, L., Kohls, K., Zeiler, S., Holz, T., Kolossa, D., 2018. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. preprint. arXiv:1808.05665.

Serrano, C.R., Sylla, P., Gao, S., Warren, M.A., 2020. Rta3: a real time adversarial attack on recurrent neural networks. In: Proceedings of the 2020 IEEE Security and Privacy Workshops (SPW), pp. 27–33.

Shi, C., Wang, Y., Chen, Y., Saxena, N., Wang, C., 2020. Wearid: low-effort wearable-assisted authentication of voice commands via cross-domain comparison without training. In: Proceedings of the Annual Computer Security Applications Conference, pp. 829–842.

Shirvanian, M., Mohammed, M., Saxena, N., Anand, S.A., 2020. Voicefox: leveraging inbuilt transcription to enhance the security of machine-human speaker verification against voice synthesis attacks. In: Proceedings of the Annual Computer Security Applications Conference, pp. 870–883.

Taori, R., Kamsetty, A., Chu, B., Vemuri, N., 2019. Targeted adversarial examples for black box audio systems. https://openreview.net/forum?id=HyGySsAct7.

Walker, P., Saxena, N., 2021. Evaluating the effectiveness of protection jamming devices in mitigating smart speaker eavesdropping attacks using gaussian white noise. In: Proceedings of the Annual Computer Security Applications Conference. Association for Computing Machinery, New York, NY, USA, pp. 414–424.

Wang, C., Anand, S.A., Liu, J., Walker, P., Chen, Y., Saxena, N., 2019. Defeating hidden audio channel attacks on voice assistants via audio-induced surface vibrations. In: Proceedings of the 35th Annual Computer Security Applications Conference. Association for Computing Machinery, New York, NY, USA, pp. 42–56.

Williams, J., Rownicka, J., 2019. Speech replay detection with x-vector attack embeddings and spectral features. preprint. arXiv:1909.10324.

Wixey, M., Johnson, S., Cristofaro, E.D., 2020. On the feasibility of acoustic attacks using commodity smart devices. In: 2020 IEEE Security and Privacy Workshops (SPW), pp. 88–97.

Yan, C., Ji, X., Wang, K., Jiang, Q., Jin, Z., Xu, W., 2022. A survey on voice assistant security: attacks and countermeasures. ACM Comput. Surv. 55, 1–36.

Yan, Q., Liu, K., Zhou, Q., Guo, H., Zhang, N., 2020. Surfingattack: interactive hidden attack on voice assistants using ultrasonic guided waves. In: Proceedings of the NDSS.

Young, J., Liao, S., Cheng, L., Hu, H., Deng, H., 2022. SkillDetective: automated policy-violation detection of voice assistant applications in the wild. In: Proceedings of the 31st USENIX Security Symposium (USENIX Security 22). USENIX Association, Boston, MA. https://www.usenix.org/conference/usenixsecurity22/presentation/young.

Zhai, T., Li, Y., Zhang, Z., Wu, B., Jiang, Y., Xia, S.T., 2021. Backdoor attack against speaker verification. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 2560–2564.

Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., Xu, W., 2017. Dolphinattack: inaudible voice commands. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. Association for Computing Machinery, New York, NY, USA, pp. 103–117.

Zhang, L., Tan, S., Wang, Z., Ren, Y., Wang, Z., Yang, J., 2020. Viblive: a continuous liveness detection for secure voice user interface in iot environment. In: Proceedings of the 2020 Annual Computer Security Applications Conference. Association for Computing Machinery, New York, NY, USA, pp. 884–896.

Zhang, N., Mi, X., Feng, X., Wang, X., Tian, Y., Qian, F., 2019. Dangerous skills: understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems. In: Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP), pp. 1381–1396.

Zhang, Z.J., Yang, E., Fang, S., 2021. Commandergabble: a universal attack against asr systems leveraging fast speech. In: Proceedings of the Annual Computer Security Applications Conference. Association for Computing Machinery, New York, NY, USA, pp. 720–731.

Zhao, Y., Togneri, R., Sreeram, V., 2018. Spoofing detection using adaptive weighting framework and clustering analysis. In: INTERSPEECH, pp. 626–630.

Jingjin Li is a second-year student at James Cook University, where she is pursuing a PhD in Cyber Security. Her research interests include security and privacy issues related to voice assistants, acoustic models, and machine learning. Jingjin completed her bachelor's degree in engineering from Harbin Institution of Technology in China, after which she pursued her master's degree in information technology from the University of Melbourne in Australia.

Under the supervision of Dr. Chao Chen, a Lecturer at RMIT University, Jingjin is conducting research to investigate the security and privacy threats associated with voice assistant applications. Her research aims to identify potential security vulnerabilities in acoustic models and machine learning algorithms used by voice assistants and to develop solutions to mitigate these threats.

Jingjin's research has significant implications for the development of voice assistant applications, as these applications continue to gain popularity in today's society. By identifying and addressing security and privacy threats, her research can help to enhance the security of voice assistant applications and protect users' personal sensitive information.