HVAC: Evading Classifier-based Defenses in Hidden Voice Attacks

Yi Wu
University of Tennessee
Knoxville, TN, USA
ywu83@vols.utk.edu

Xiangyu Xu
Shanghai Jiao Tong University
Shanghai, China
chillex@sjtu.edu.cn

Payton Walker
University of Alabama at Birmingham
Birmingham, AL, USA
prw0007@uab.edu

Jian Liu
University of Tennesse
Knoxville, TN, USA
jliu@utk.edu

Nitesh Saxena
University of Alabama at Birmingham
Birmingham, AL, USA
saxena@uab.edu

Yingying Chen
Rutgers University
New Brunswick, NJ, USA
yingche@scarletmail.rutgers.edu

Jiadi Yu
Shanghai Jiao Tong University
Shanghai, China
jiadiyu@sjtu.edu.cn

ABSTRACT

Recent years have witnessed the rapid development of automatic speech recognition (ASR) systems, providing a practical voice-user interface for widely deployed smart devices. With the ever-growing deployment of such an interface, several voice-based attack schemes have been proposed towards current ASR systems to exploit certain vulnerabilities. Posing one of the more serious threats, *hidden voice attack* uses the human-machine perception gap to generate obfuscated/hidden voice commands that are unintelligible to human listeners but can be interpreted as commands by machines. However, due to the nature of hidden voice commands (i.e., normal and obfuscated samples exhibit a significant difference in their acoustic features), recent studies show that they can be easily detected and defended by a pre-trained classifier, thereby making it less threatening. In this paper, we validate that such a defense strategy can be circumvented with a more advanced type of hidden voice attack called *HVAC*[1]. Our proposed HVAC attack can easily bypass the existing learning-based defense classifiers while preserving all the essential characteristics of hidden voice attacks (i.e., unintelligible to humans and recognizable to machines). Specifically, we find that all classifier-based defenses build on top of classification models that are trained with acoustic features extracted from the entire audio of normal and obfuscated samples. However, only speech parts (i.e., human voice parts) of these samples contain the useful linguistic information needed for machine transcription. We thus propose a fusion-based method to combine the normal sample and corresponding obfuscated sample as a hybrid HVAC command, which can effectively cheat the defense classifiers. Moreover, to make the command more unintelligible to humans, we tune the speed and pitch of the sample and make it even more distorted in the time domain while ensuring it can still be recognized by machines. Extensive physical over-the-air experiments demonstrate the robustness and generalizability of our HVAC attack under different realistic attack scenarios. Results show that our HVAC commands can achieve an average 94.1% success rate of bypassing machine-learning-based defense approaches under various realistic settings.

CCS CONCEPTS

• **Security and privacy → Systems security**.

KEYWORDS

voice control, hidden voice command, classifier-based defense

ACM Reference Format:
Yi Wu, Xiangyu Xu, Payton Walker, Jian Liu, Nitesh Saxena, Yingying Chen, and Jiadi Yu. 2021. HVAC: Evading Classifier-based Defenses in Hidden Voice Attacks. In *2021 ACM Asia Conference on Computer and Communications Security (ASIA CCS '21), June 7–11, 2021, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3433210.3437523

1 INTRODUCTION

Driven by the rapid development of voice-user interfaces, automatic speech recognition (ASR) systems have been implemented in many facets of our daily lives. They have been integrated into smartphones (e.g., Apple Siri [5]), smart speakers (e.g., Amazon Alexa [4], Google Home [15]), smart home appliances (e.g., smart TVs), and even vehicles (e.g., Tesla Voice Commands [34]). With these ubiquitous voice-user interfaces, people can just speak to their devices to make phone calls, set up timers & alarm clocks, and control IoT devices, etc. Given higher permission, voice assistants can even unlock the doors of a car [34], make mobile payments [17], and schedule personal appointments [24]. A market share report

[1]HVAC denotes "**H**idden **V**oice **A**ttacks in presence of **C**lassifier"

ASIA CCS '21, June 7–11, 2021, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-8287-8/21/06…$15.00
https://doi.org/10.1145/3433210.3437523

suggests that there are an estimated 3.25 billion digital voice assistants being used in devices around the world in 2019, and the number would reach around eight billion units by 2023 [10].

Unfortunately, despite the great convenience provided by these voice assistants, many existing studies have demonstrated that they are vulnerable to various kinds of voice-based attacks. For instance, deep neural networks (DNNs), serving as the computation cores of ASR systems, have been proved to be fragile against *adversarial machine learning attacks* [7–9, 29, 33, 37], where an adversary could inject an imperceptible perturbation into a normal audio command to make recognition models output any adversary-desired transcription. Another type of attack is *hidden voice attack* [1, 6, 35]. Different from adversarial machine learning attacks, hidden voice attacks use a different attacking angle: an adversary could generate a noise-like obfuscated audio sample (i.e., a hidden voice command), which sounds unintelligible to humans but can be correctly recognized as the target transcription by ASR systems. To generate such a hidden voice command, the adversary could use the human-machine perception gap to either adjust the parameters of Mel frequency cepstral coefficients (MFCCs) to produce lower fidelity (to humans) but recognizable (to machine) audio samples [6, 35] or randomize the time-domain signal of normal commands while ensuring the frequency-domain is unchanged [1].

Compared with adversarial machine learning attacks, hidden voice attacks are more realistic and present a relatively better performance in practice, since the obfuscated/hidden audio samples are less sensitive to the audio distortions introduced in over-the-air transmissions than imperceptible perturbations. This makes them easier and more effective to launch in practice, leading to more severe security concerns. For instance, hidden voice commands could be stealthily embedded in a broadcast or a trending Youtube video, potentially having a broader attack range and infecting a large number of unnoticed victims [6]. The adversary could also play the hidden voice command near the victim ASR device to make the device execute the obfuscated malicious command without causing any suspicion. Therefore, the central focus of this paper is on hidden voice attacks.

Classifier-based Defenses Against Hidden Voice Attacks. Although hidden voice attacks introduce a severe threat against voice assistants, existing studies have demonstrated that they can be easily detected and defended using a machine-learning-based classifier [6]. Since hidden voice commands sound like noises, the hidden voice samples exhibit significantly different acoustic characteristics compared with the normal human voice. Thus, devices with ASR system could incorporate a machine learning-based classifier as a verification step, classifying the received speech as being issued by a human or machine-generated. Specifically, through training a classifier using the acoustic features (e.g., zero-crossing rate, MFCC coefficients, and chroma vectors) extracted from two types of audio samples, a simple logistic regression classifier could distinguish hidden voice commands from normal commands with a 99.8% true positive rate and a 0.2% false-positive rate [6]. As shown in Figure 1, without the protection from a defense classifier, the hidden voice command can be directly received and transcribed by the victim ASR system. However, in the presence of a defense classifier, the hidden voice command will be correctly detected, thereby being rejected by the ASR system. Given the existence of classifier-based

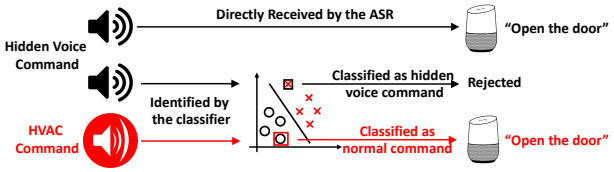


Figure 1: HVAC commands vs. hidden voice commands in the presence of a defense classifier.

defenses, we take one step further and explore: *How hard to defeat the classifier-based defenses and launch an advanced hidden voice attack to bypass the classifier-based defenses?*

Evading Classifier-based Defenses. Existing learning-based defense classifiers tend to utilize the acoustic features of the entire audio of normal and obfuscated/hidden samples, while only the features of their speech parts (i.e., human-voice parts) contain useful linguistic information for translation to ASR systems. Relying on this observation, in this paper, we propose *HVAC (Hidden Voice Attacks in presence of Classifier)*, a more powerful hidden voice attack that can circumvent existing learning-based defense classifiers while preserving all essential characteristics of hidden voice attacks (i.e., unintelligible to humans but recognizable to ASR systems). To make our attack more realistic, HVAC commands are generated in a black-box setting, where the adversary has no internal knowledge of the target ASR system (e.g., model architecture, parameters). Specifically, to create such an HVAC command, we first generate a hidden voice command through tuning MFCC parameters to produce obfuscated audio, which is the similar black-box method of the prior work [6]. Through elaborately mixing the signals of the hidden voice command and its corresponding normal command, the adversary could push the acoustic features of the fused HVAC commands towards the features of the normal human voice, potentially making the defense classifier make false predictions, as shown in Figure 1. Furthermore, to make the generated HVAC command more unintelligible to humans, we tune the speed and pitch of the command and make the sample more distorted in the time domain while ensuring it can still be recognized by the ASR system. Our main contributions are summarized as follows:

- We dissect existing defense techniques and find these classifier-based strategies are not robust enough to defend against hidden voice attacks. The effectiveness of these defense classifiers largely relies on the acoustic features extracted from normal and obfuscated samples, an adversary thus can make them yield false predictions through modifying the hidden command's features while preserving the properties of hidden voice attacks.
- We present our novel design for an HVAC (Hidden Voice Attacks in presence of Classifier) attack against ASR systems that can bypass the state-of-the-art classifier-based defense approaches. The generated HVAC commands could be successfully recognized by ASR systems while remaining indecipherable to human listeners similar to existing hidden voice attacks (in the absence of classifier defenses).
- To further improve the unintelligibly of the generated HVAC commands to humans, in the generation phase we also propose to

employ a human perception mask, including audio speed tuning, audio pitch tuning, and time-domain inversion.
- Extensive physical over-the-air experiments demonstrated the robustness and generalizability of our HVAC commands under various realistic attack settings (e.g., in different environments, with various commercial voice assistants, and attack distances). The results showed the effectiveness of our proposed HVAC attack with an average 86.1% ASR transcription recognition accuracy and a high success rate of bypassing various defense classifiers of 94.1%.

2 STUDY OF EXISTING HIDDEN VOICE COMMANDS: ATTACK AND DEFENSE

In this section, we first give a brief introduction on the background of automatic speech recognition (ASR) system. Then, we present a comprehensive study on the prior work [6] (i.e., the first work of hidden voice attacks) by reproducing its corresponding attack and defense strategies, respectively.

2.1 Background of Automatic Speech Recognition System

An automatic speech recognition (ASR) system is a voice processing procedure that automatically converts human speech input to text output. As shown in Figure 2, a typical speech recognition system takes four steps (i.e., pre-processing, feature extraction, model-based prediction, and post-processing) to generate the speech recognition results from the received human speech. As the first step, pre-processing mainly focuses on filtering out background noises as well as frequencies that are outside the range of human speech. Then in the feature extraction step, signal processing algorithms are applied to capture effective audio features from the denoised signal. Among current ASR systems, the most wildly applied features in this step are Mel Frequency Cepstral Coefficient (MFCC) features, which can capture phonetic characteristics of human speech [20, 26]. After that, the extracted features are fed into a pre-deployed model to generate text predictions. Either a statistical inference-based or machine-learning-based model can be applied in this step. And for the final step, post-processing votes or ranks the generated text predictions to decide the text output as the result of speech recognition.

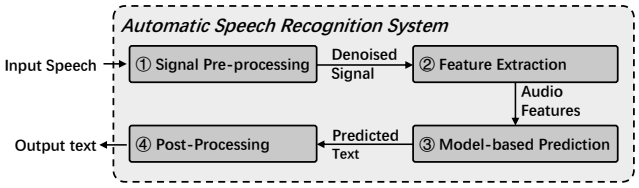


Figure 2: Illustration of a typical speech recognition system.

2.2 Performance Study of Existing Hidden Voice Command Attack

As the pioneering work of hidden voice attacks, Carlini *et al.* [6] have shown that an adversary could generate *hidden voice commands*, which are intelligible to ASR systems but not to humans,

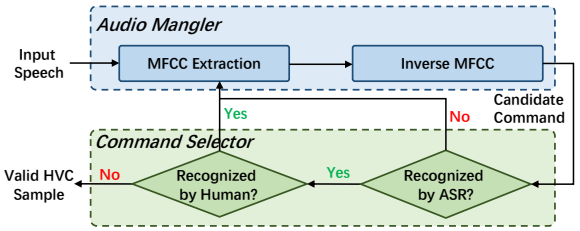


Figure 3: Attack flow of hidden voice command [6].

Table 1: Performance comparison of existing hidden voice commands and our implementation.

	Results in Hidden Voice Command [6]			Our Implementation Results
Commands	"OK Google"	"Turn on airplane mode"	"Call 911"	500 commands in the telephone corpus [23]
Normal	90% (36/40)	75% (30/40)	90% (36/40)	83% (415/500)
Hidden Voice Command	95% (38/40)	45% (18/40)	40% (16/40)	56% (279/500)

in a black-box attack setting, where the adversary has no internal knowledge of the target ASR system.

Essentially, a hidden voice command is generated from a normal speech signal by preserving most of the MFCC features. Figure 3 illustrates the workflow for producing a valid hidden voice command. After a normal command is produced by the adversary, it is sent to the *audio mangler*, where the MFCC features of the given command are extracted. Specifically, 5 MFCC parameters (i.e., *wintime, hoptime, numcep, nbands*, and *maxfreq* listed in Table 3) pre-determined by the attacker are isolated during the MFCC feature extraction. The choices of these parameters determine the dimension & resolution of the extracted MFCC features. Then, the extracted MFCC features are converted back to an audio sample through inverse MFCC operation, which produces a candidate command sample. Through the MFCC extraction-inverse MFCC process, the candidate command sample shares the same MFCC features with the normal command, which will tend to make ASR systems output the same transcription. Meanwhile, it only preserves features that are useful for machine understanding, while features contributing to human perception are discarded. Then, the candidate command sample is sent to the *Command Selector* shown in Figure 3, where both ASR system recognition test and human perception test are applied to the candidate sample. If the candidate sample can be recognized by the given ASR system correctly, and also cannot be understood by humans at the same time, then the command is a valid hidden voice command. Otherwise, if the candidate command fails either the ASR system recognition test or human perception test, it needs to be reproduced by adjusting the 5 MFCC parameters in audio mangler, until it can pass the two tests in the command selector. After all the processes, a valid hidden voice command is generated as a black-box attack towards ASR systems.

Since the prior work [6] only reported the performance of three specific commands, i.e., "OK Google", "Call 911" and "Turn on airplane mode", under black-box attack scenario, to validate the generality and robustness of the hidden voice attack, we reproduced the procedure in Figure 2 to generate and test hidden voice commands on a larger set of samples. In particular, 500 hidden voice

Table 2: Performance comparison of the defense classifier proposed in hidden voice commands [6] and our reproduction.

	Hidden Voice Command [6] Loigistic Regression	Reproduced (SVM)	Reproduced (Logistic Regression)
Attack Detection Rate	99.80%	99.93%	99.90%

commands were generated from telephone corpus [22]. For the machine understanding test, the experiment was conducted in a small office (approximately 4.5$sqft$), with an external microphone (iTalk-02) to record the attack command while the environment sound level was 55dB according to the noise meter measurement. A loud stereo speaker (Logitech Z623) was set up for playing the attack commands and placed 1.5m away from the smartphone. The target ASR system is the Google speech recognition system with open APIs.

Table 1 shows the performance of hidden voice command attack, both from the prior work [6] and our reproduction. It can be seen that for machine understanding, 56% of the generated hidden voice commands could be understood by the Google speech recognition system (given the baseline as 83% of the normal commands can be recognized). The reproduced results show a comparable level to the prior work [6], validating the effectiveness of this way to generate valid hidden voice commands.

2.3 Performance Study of Learning-based Defense Classifier

Although hidden voice command attack seems creating a great threat to ASR systems as a black-box attack, there have already been some defense strategies. The notification and authentication mechanisms of ASR systems could potentially prevent such hidden voice command attack. For instance, ASR systems may notify users via lights and/or beeps, or require audio CAPTCHA [25] whenever sensitive operations are requested by users. Speaker identification and verification techniques [30, 31] are also effective in defending against hidden voice command attacks as they require user authentication before executing the speech command.

Most importantly, as pointed out in the prior work [6], training a simple classifier (i.e., logistic regression) with the acoustic features extracted from hidden voice command and normal speech command could almost fully defend against hidden voice command attack. In other words, once the attack strategy is known, the hidden voice command attack can be easily defended by integrating a simple defense classifier in ASR system, which greatly reduces the threat of hidden voice command attacks.

To validate the effectiveness of the learning-based defense classifier against hidden voice command attacks, we implemented two classifiers using logistic regression and SVM respectively. The models were trained with statistical features from partial Vystadial project [23] including 20,000 samples of normal human speech and corresponding 20,000 samples of hidden voice commands. The statistical features were extracted from the audio signal using *pyAudioAnalysis* [13] which splits the input audio signal into short-term windows (frames) and computes a total of 68 features (e.g., MFCC coefficients, chroma vectors, spectral spread) for each frame. The mean and standard deviation of these features among all frames make up a total of 136 features representing the audio signal.

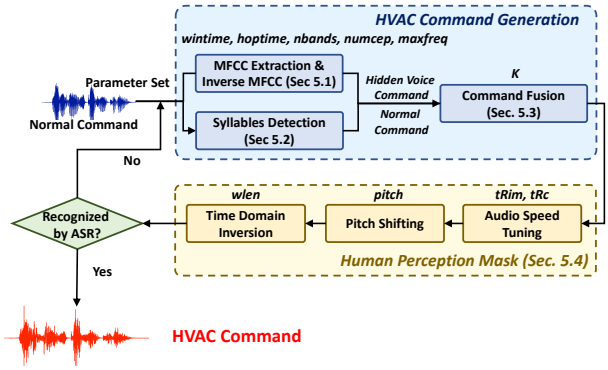


Figure 4: HVAC Attack Overview.

Table 2 shows the performance of learning-based defense classifiers, both from the prior work [6] and our reproduction with SVM and logistic regression, respectively. All three results are close to 100% attack detection rate, showing that the learning-based defense classifier is extremely effective to detect and defend against hidden voice commands.

3 HVAC: THREAT MODEL & ATTACK OVERVIEW

In this section, we will introduce the threat model and the overall attack flow of our proposed HVAC attack.

3.1 Threat Model

Considering the weakness of the prior work [6], we present a stronger threat model. The goal of the adversary is to launch an attack targeting ASR systems, by generating commands that can be recognized by the targeted ASR system but cannot be understood by humans. Moreover, the generated commands could also pass the defense classifier that is trained with hidden voice commands. The specific assumptions from the adversary's perspective are listed below.

- No prior knowledge of the target ASR system is required for the adversary to launch the attack, which makes it a black-box attack.
- The targeted ASR system could apply a training-based classifier (e.g., Logistic Regression, Support Vector Machine, Deep Neural Networks) to defend against hidden voice attacks.
- The adversary could get access to the same or similar ASR system (as a black-box) to test if the generated commands can be correctly recognized by the ASR system.

As the proposed HVAC attack is built in a black-box setting, the attack is widely applicable to attacking various ASR systems including those commercial closed sourced speech recognition devices (e.g., Google Home and Alexa Echo). Some representative attack scenarios include embedding the generated hidden command into the audio tracks of regular media (e.g., Youtube videos and radios) to control all the ASR devices exposed to that media, and playing

Table 3: Parameter Set

Type	Parameter	Description
MFCC Parameters	*wintime*	Sliding window size for MFCC extraction
	hoptime	Step between successive windows
	numcep	Number of cepstras to return
	nbands	Number of warped spectral bands to use
	maxfreq	The highest band edge of mel filters
Command Fusion	*K*	The hybrid fusion ratio
Human Perception Mask Parameters	*tRim*	Speed tuning ratio for the speech part
	tRc	Speed tuning ratio for the non-speech part
	pitch	Pitch shifting ratio
	wlen	Sliding window size for time domain inversion

the designed attack sample in the physical proximity of the victim ASR devices, etc.

3.2 Attack Overview

The goal of our attack is to generate an HVAC command which can: 1) be correctly recognized by ASR systems; 2) bypass classifier-based defenses; and 3) maintain unintelligible to humans.

To achieve the above attack goal, we propose a generic audio-processing flow that can generate HVAC commands from any normal commands, as shown in Figure 4. There are two main components including *HVAC Command Generation* which generates HVAC commands that can bypass classifier-based defenses while still being correctly recognized by ASR systems, and *Human Perception Mask* which is used to further enhance the unintelligibility of the generated commands. Specifically, an understandable normal command (adversary-desired command) needs to be produced first. For instance, an adversary could either use his pre-recorded voice command or create such a normal command using a text-to-speech (TTS) engine. Next, the adversary generates the corresponding *hidden voice command* through tuning MFCC parameters during the process of MFCC extraction and inverse MFCC, which is similar to the black-box method proposed by Carlini *et al.* [6]. To be more specific, the adversary first extracts commonly used acoustic features (i.e., MFCC) from the normal command, and then performs inverse MFCC to convert the extracted MFCC features back to an audio sample. In this step, 5 MFCC parameters listed in Table 3 are used to regulate the granularity of the extracted features, including sliding window size, step between successive windows, number of cepstras to return, number of warped spectral bands to use, and the highest band edge of mel filters. These parameters are initialized with random values at the beginning and then iteratively adjusted according to whether the command can be recognized by ASR systems. After the step of inverse MFCC, the generated audio sample only contains MFCC features which are necessary to the machine-based speech recognition while disregarding other useful features for human comprehension. This essentially makes the command recognizable to machines but unintelligible to humans.

Meanwhile, a syllables detection algorithm [19] is implemented to detect the speech parts and non-speech parts of the normal command. The adversary then fuses the normal command and the hidden voice command in Command Fusion, where the non-speech parts are directly copied from the normal command and the speech parts are a combination of the two audio samples. Through command fusion, parts of the HVAC command preserve the characteristics of the normal command, which makes traditional classifier-based defenses that extract features from the whole audio hard to distinguish it from normal commands. The parameter K represents the proportion of the hidden voice command in the HVAC command. A larger K indicates more features from the hidden voice command are preserved, and vice versa. Same with the 5 MFCC parameters, K is also initialized with a random number at the beginning of command generation.

To further downgrade human intelligibility to the generated command, the adversary would tune the speed & pitch of the HVAC command and inverse it in the time domain to make it distorted and lose consistency. The four human perception mask parameters (i.e., *tRim, tRc, pitch,* and *wlen*) control the scale of the audio distortion. A more distorted HVAC command would be harder for humans to understand, but it may also let ASR systems couldn't recognize it. Therefore, the adversary needs to test whether the HVAC command can be recognized by the target ASR system. If the command is not recognizable, the adversary needs to start over the whole process using a new set of parameters. It's essential for the adversary to try different sets of parameters and find an appropriate trade-off between machine understanding and human intelligibility.

4 ATTACK DESIGN

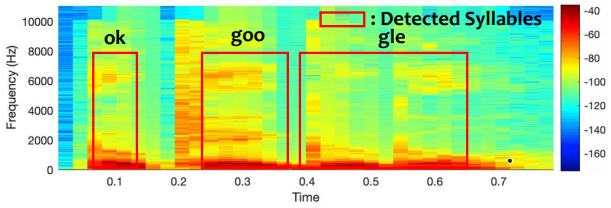
In this section, we will introduce the core components of generating an HVAC command from a normal command.

4.1 MFCC Extraction & Inversion

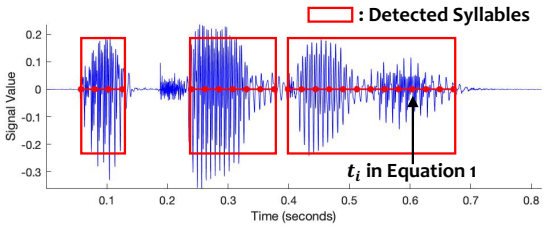
Similar to the black-box attack method in the prior work [6], the adversary first extracts MFCC features from the normal command and then inverses these features back to an audio sample, which is the corresponding hidden voice command. The obfuscated hidden voice command thus shares the same MFCC features with the normal command, which will tend to make ASR systems output the same transcription. However, it only preserves features that are useful for machine understanding, while features contributing to human perception are discarded. The dimension and resolution of the extracted MFCC features are determined by the 5 MFCC parameters (i.e., *wintime, hoptime, numcep, nbands*, and *maxfreq* listed in Table 3), which further influences the sound quality of the re-constructed hidden voice command. A higher dimension & resolution of the extracted MFCC features would help the ASR system understand the hidden voice command more precisely, but also makes it more perceptible to humans. If the reconstructed audio can be easily understood by humans, the adversary can generate a new candidate by adjusting these MFCC parameters to produce lower fidelity audio.

4.2 Syllables Detection

For a command or a short sentence, it contains speech parts (presence of human speech) and non-speech parts (absence of human speech). To ASR systems, only speech parts contain the information that is useful for speech recognition. However, classifier-based defenses usually extract acoustic features from the entire audio sample (both speech and non-speech parts) to build learning-based classifiers. To make the HVAC command unintelligible to humans while being able to bypass classifier-based defenses, the adversary



(a) Audio signal and the detected syllables in the frequency domain after STFT.



(b) Audio signal and the detected syllables in the time domain.

Figure 5: Illustration of Syllables Detection.

needs to combine these two parts of hidden commands and normal commands in a way that makes the acoustic features extracted from HVAC commands close to the features extracted from normal commands. Thus, the adversary first needs to detect and segment speech parts and non-speech parts of the normal command.

The speech parts consist of multiple audio segments, and each segment represents a single syllable. Specifically, the adversary uses the syllable segmentation algorithm proposed by Harma *et al.* [19], in which the key idea is to find segments in the audio sample which contains a relatively larger power density than a pre-defined threshold. To be more specific, the adversary first calculates the Short-time Fourier transform (STFT) of the input normal command sample, which is represented as a matrix $S(f,t)$, where f is frequency and t is time. A minimum magnitude threshold τ is set for the syllables. The adversary then repeats the following steps: for the *nth* syllable, the adversary first finds f_n and t_n where $S(f_n,t_n)$ is the maximum value in the spectrogram. The adversary stops searching if $S(f_n,t_n)$ is smaller than τ. The adversary then finds the starting point and ending point of the syllable segment t_s and t_e through solving the following equation:

$$\begin{aligned} &\arg\max_{t_s,t_e} t_e - t_s, \\ &s.t., t_s < t_n, t_e > t_n, \\ &s.t., \forall t_i \in [t_s,t_e],\ max(S(f,t_i)) > S(f_n,t_n) - \tau. \end{aligned} \tag{1}$$

After finding the starting and ending points of a syllable, the adversary sets $S(f,[t_s,t_e])$ to 0 and search for the maximum value (i.e., $S(f_n,t_n)$) in the spectrogram again for the next syllable. Figure 5 represents an example of implementing the syllables detection algorithm on a normal command with "*Ok Google*" as its corresponding transcription. Figure 5 (a) and Figure 5 (b) represent the audio signal in frequency and time domain, respectively. The detected syllables are labeled with red rectangles and t_i (used for searching for t_e and t_s) is marked with red dots. We can clearly observe that

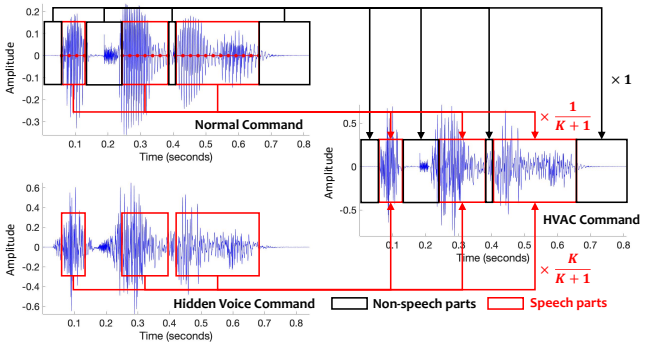


Figure 6: Illustration of fusing normal command and hidden voice command.

the detected syllables segments contain relatively larger energy and match the voice activity in the audio sample.

4.3 HVAC Command Generation

After the detection of syllables, the adversary fuses the normal command and hidden voice command together to generate the HVAC command. The fused HVAC command thus carries the characteristics from both audio samples, making it hard to be understood by humans while exhibits more similar acoustic features to the normal command. This will help the fused command have higher chance to bypass learning-based defense classifiers. Each detected syllable would be regarded as a segment of the speech parts of the normal command. Since the hidden voice command shares the same length and is synchronized with the normal command, the adversary can directly combine these two samples. Specifically, as shown in Figure 6, the non-speech parts of the HVAC command are directly copied from the normal command without any modification, while the speech parts are the combination of both audio samples. Since the non-speech parts won't affect machine understanding and human perception, preserving the whole non-speech parts from the normal command would maximize the proportion of the characteristics inherent from it in the HVAC command. This further makes the HVAC command exhibit a similar pattern with the normal command, enabling it to bypass the existing learning-based defense classifier. To be more specific, given the normal command as N and the hidden voice command as H, the fused HVAC command F can be derived as:

$$F(t) = \begin{cases} \dfrac{N(t) + H(t) \times K}{K+1}, & t \in speech\ parts, \\ N(t), & t \in non-speech\ parts, \end{cases} \tag{2}$$

where K is the pre-defined hybrid fusion ratio, which represents the proportion of the hidden voice command in the HVAC command, ranging from 2 to infinite.

To demonstrate the fused HVAC commands are indeed hard to be distinguished from normal commands, we generate 200 normal commands, 200 hidden voice commands and 200 HVAC commands. For each command sample, we extract 136 acoustic features as aforementioned in Section 2.3. For visualization purposes, we use three representative features to differentiate the three types of commands, as shown in Figure 7. Specifically, the features we visualized

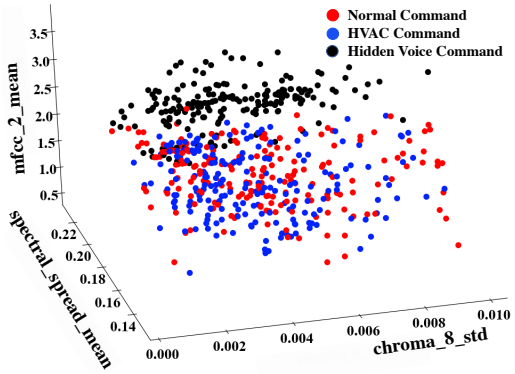


Figure 7: Illustration of distinguishing hidden voice commands from normal & HVAC commands.

are the mean of the second MFCC coefficient, the mean of the second central moment of the spectrum, and the standard deviation of the 8th chroma vector. We can observe that it is very hard to distinguish HVAC commands from normal commands since most of these two types of samples are mixed with each other, while the hidden voice commands are clustered separately from them. When training the defense classifier using the complete 136 features, we observe that nearly all HVAC commands would be classified as normal commands, which will be demonstrated with more details in Section 5.

4.4 Human Perception Mask

To further increase the unintelligibility of the generated HVAC sample, we propose a human perception mask to make the audio sample even more distorted in the time domain while ensuring it can still be recognized by machines.

Audio Speed Tuning. When the same speech is spoken at a faster rate, it would be harder for humans to understand [18]. However, although the duration of the command is shorter after speeding up, the frequency spectrum will remain mostly the same and won't largely affect the performance of being recognized by ASR systems [1]. The adversary thus can further downgrade human perception through speeding up the HVAC command. Different from existing studies where the whole audio is sped up with a constant ratio, in our attack, the adversary speeds up the speech parts and non-speech parts of the HVAC command with two different ratios, *tRim* and *tRc*, respectively. Compared with *tRim*, *tRc* is slightly smaller. This would still make the command harder for humans to perceive but increase the probability of being correctly transcribed by the ASR system. This is because when the ratio is too high, the ASR system will likely miss words or treat two separate words as a single one. Slowing down the ratio for non-speech parts a little would enlarge the gap between two successive syllables therefore largely decreasing the probability of the occurrence of this type of mistake. Meanwhile, the transcription is still spoken at a faster rate which makes it harder for humans to understand. Moreover, this would also disturb the consistency of features extracted from the whole audio which is used by current learning-based defense approaches.

Pitch Shifting. In addition to speeding up the HVAC command, the adversary can also scale up the pitch. Scaling up the pitch would make the command shriller and less likely to be regarded as a human speech. The adversary uses the pitch shifting algorithm provided by the Time-scale modification (TSM) Toolbox [11] in Matlab. Given the pre-defined pitch-shifting ratio *pitch* in the parameter set, the algorithm will change the pitch of the signal by $\frac{pitch}{100}$ semitones while ensuring the original sampling rate and audio speed is not influenced. Specifically, given an input audio whose sampling rate is f_{old}, the algorithm first resamples it to a new frequency f_{new}. When the resampled signal is played back with f_{old}, the pitch would change by $log_{2^{\frac{1}{12}}} \frac{f_{old}}{f_{new}}$ (equals to $\frac{pitch}{100}$), and the length of it would change by a factor of $\frac{f_{new}}{f_{old}}$. Therefore, another TSM algorithm that rescales the time-axis of the resampled audio is further implemented to compensate for the reduced length while ensuring the original audio speed and the modified pitch won't be influenced when playing back using the original sampling rate f_{old}.

Time Domain Inversion. Proposed by Abdullah *et al.* [1], time-domain inversion is a sliding-window based algorithm which inverses the audio signal in time domain in each window. There is no overlap between successive windows and the size of the windows is pre-defined in the parameter set as *wlen*. This would preserve the frequency characteristics in each window, but make the audio signal discontinuous and perturbed in the time domain which makes it harder for humans to understand.

The whole process of the human perception mask is illustrated in Figure 8. For visualization purposes, we only show the effect on a small segment (i.e., approximately 0.075 seconds) of the HVAC command audio sample. Specifically, the speed-up ratio is 1.25 for both speech parts and non-speech parts, the pitch tuning ratio is 200, and the sliding window size *wlen* for time domain inversion is 10. We can see after adding the human perception mask, the HVAC command is further distorted, making it even harder for humans to understand.

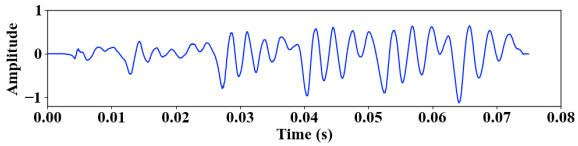
The adversary finally feeds the HVAC command after adding the human perception mask into the target ASR system and test whether it could be successfully transcribed. If it could not be recognized, the adversary will start over the whole process using a new set of parameters. The adversary could try to increase the dimension of the MFCC features (i.e., increase *numcep*), reduce the characteristics inherent from the hidden voice command (i.e., reduce K), or make the HVAC command less distorted (i.e., decrease *wlen*), etc.

5 EVALUATION

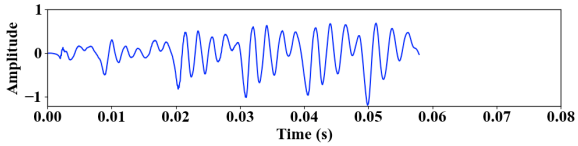
In this section, we first present the experimental methodology and then evaluate the performance of HVAC commands with respect to their intelligibility level, probability of bypassing defense classifiers, and the ASR recognition accuracy under over-the-air transmission.

5.1 Experimental Methodology

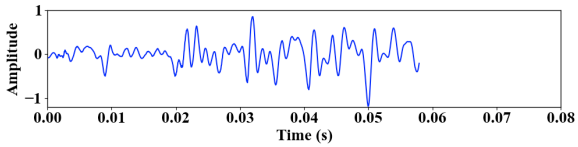
Data Collection & Experimental Setup. We select 15 speech commands for evaluation as shown in Table 6 in Appendix, which exhibit a series of potential practical threats. The normal command version of these 15 speech commands is generated through the IBM Watson text to speech service APIs [28]. After that, with the generated normal commands as input, we apply the framework



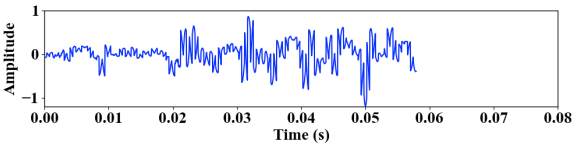
(a) HVAC command segment



(b) After speed tuning



(c) After pitch tuning



(d) After time domain inversion

Figure 8: Illustration of Human Perception Mask.

illustrated in Figure 4 to generate corresponding HVAC commands. Specifically, during the HVAC command generation, the Google speech recognition system API [16] is leveraged to test whether the generated digital command can be recognized by ASR systems.

We consider the most realistic attack scenarios where the HVAC commands are played over-the-air to attack the ASR systems. In particular, the HVAC commands are played in three different indoor setups using two models of loudspeakers and received by three different voice assistants. As illustrated in Figure 9 (a), the first setup takes place in a bedroom, with an Amazon Echo acting as the victim voice assistant. The HVAC commands are played using a Logitech Z623 loudspeaker which is placed one meter away from the Amazon Echo. Similarly, setup 2 shown in Figure 9 (b) is also a bedroom setup. A Google Home acts as the victim ASR system and the HVAC commands are played using an Edifier R980T loudspeaker [12] placed one meter away. Lastly, an office setup is shown in Figure 9 (c) as setup 3. An iTalk-02 360-degree omnidirectional microphone is utilized to emulate the builtin microphones of smart home appliances and is placed 1 to 3 meters away from the Edifier R980T loudspeaker. In each setup, we played the generated HVAC commands 20 times, resulting a total of 2,100 HVAC command recordings.

To validate that our HVAC commands can bypass learning-based classifiers, we use two sets of data to train the classifiers. First, we train the classifiers using normal commands and hidden voice commands proposed in the prior work [6], each with 20,000 training samples. The normal commands are randomly chosen from Vystadial project [23], and we generate their corresponding hidden voice commands and HVAC commands. In particular, statistical features are extracted from the command samples using *pyAudioAnalysis* [13]. We use four machine learning algorithms (i.e., logistic regression, SVM, random forest and gradient boosting) to train the classifiers leveraging *sklearn* open source library [27]. Second, we aim to build more advanced defense classifiers by involving HVAC commands. The training process is almost the same except the training set consists of normal commands and HVAC commands, each with 20,000 samples.

Evaluation Metrics. We evaluate our HVAC commands using three different metrics:

- *Normalized Mel Cepstral Distortion (MCD)*: The Mel cepstral distortion is used to evaluate the human intelligibility level of HVAC commands, defined as

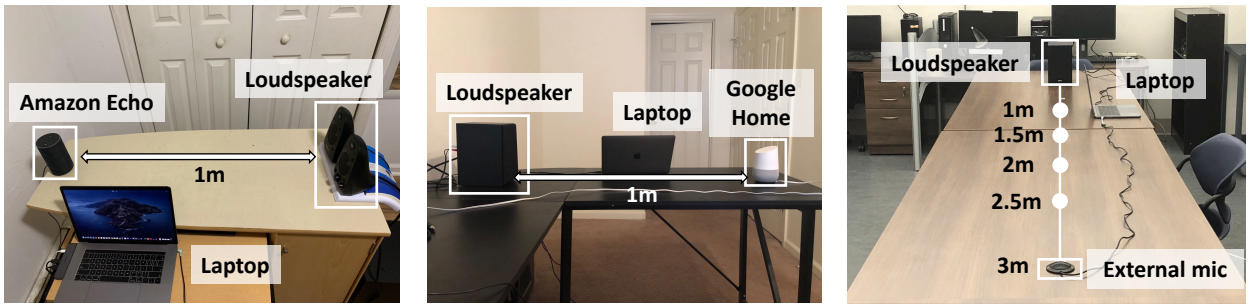
$$MCD = (10/ln(10)) \times \sqrt{2 \times \sum_{d=1}^{24} (mc_n^d - mc_h^d)^2},$$

  where mc_n^d and mc_h^d are the dth mel-cepstra of two audio signal, respectively. To generalize the results, the MCD is normalized using $\frac{MCD}{len(sequence)}$.
- *ASR Recognition Accuracy*: We leverage the Levenshtein distance ratio to measure the recognition accuracy of ASR systems. Given the ground truth text of an HVAC command as X, and the corresponding recognition result of the ASR system as Y, then the Levenshtein distance ratio is defined as $R = \frac{len(X)-leven(X,Y)}{len(X)}$, where $leven(X, Y)$ is the Levenshtein distance between X and Y, and $len(X)$ is the string length of text X.
- *Success Rate of Bypassing Defenses*: The possibility that an HVAC command is recognized as a normal command by the defense classifier, which is defined as the ratio of HVAC commands that are not correctly recognized out of all the HVAC commands.

5.2 Intelligibility Test

We use the normalized Mel Cepstral Distortion (MCD) to evaluate the human intelligibility level of our HVAC commands compared with hidden voice commands [6]. The normalized MCD between two audio samples indicates the level of distortion needed to transfer from one to another [8]. If the MCD between the HVAC command and the normal command is larger than the MCD between the hidden voice command and the normal command, it means that the HVAC command shares a lower similarity with the normal command in phonetic characteristics, indicating it's more distorted and less likely to be recognized by human listeners.

Although some studies perform a user study (e.g., Amazon Turk Study) to test human intelligibility [6], we do not apply this method for the following reasons: 1) As pointed out by Abdullah *et al.* [1], this type of user study has many uncontrollable variables. For instance, the participants could have different ages, first languages, listening equipment (e.g., a headphone or a loudspeaker). These will all affect their perception of the HVAC commands, which may lead to significantly biased results. 2) When the participants are listening to our HVAC commands, they would be aware that there's a transcription hidden in it, and would thus be more focused on revealing the hidden transcription instead of regarding the command as meaningless noise. However, when the attack is launched in practice, nobody would try to perceive the potential transcriptions



(a) Setup1: Attack setting for Amazon Echo (b) Setup2: Attack setting for Google Home (c) Setup3: Attack setting for External Microphone

Figure 9: Experimental setups.

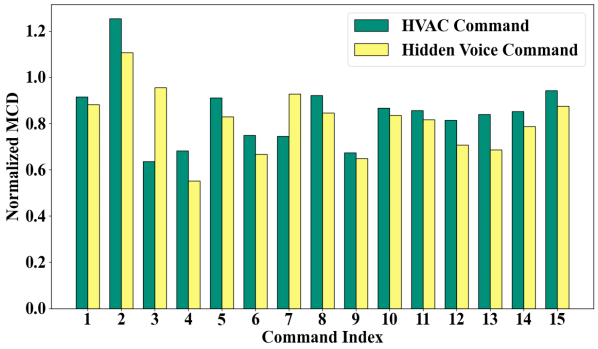


Figure 10: Intelligibility test of HVAC commands compared with hidden voice command [6].

hidden in a noise-like sound. Therefore, we do not feel a live user study is the appropriate method for determining the intelligibility level of our HVAC commands.

For each generated hidden voice command [6] and HVAC command, we calculate the normalized MCD between these hidden commands and their corresponding normal commands. The experimental results are shown in Figure 10. It's clear that our HVAC commands have comparable distortion distance to normal commands compared with hidden voice commands [6], with 13 out of 15 are even more distorted. These results demonstrate that our HVAC commands have comparable unintelligibility to humans as the prior work [6].

5.3 Validation of Bypassing Defense Classifier

We first test whether the HVAC commands could bypass defense classifiers trained using normal commands and hidden voice commands [6]. Specifically, four machine learning algorithms including logistic regression, SVM, random forest and gradient boosting are utilized to train these classifiers. We use 1,000 recorded HVAC command samples collected by the external microphone in setup 3 as our testing set. The success rate of bypassing defense classifiers of these samples is shown in Table 4. We find that nearly all of our HVAC commands can bypass the existing state-of-the-art classifiers. When the classifier uses Random Forest, we can maximize the success rate of bypassing defense classifiers to 100%. Among

Table 4: Success rate of bypassing defense classifiers (trained using normal commands and hidden voice commands [6]).

Classifier Name	Logistic Regression	SVM	Random Forest	Gradient Boosting
Success Rate of Bypassing Defenses	90.4%	98.8%	100.0%	99.7%

Table 5: Success rate of bypassing defense classifiers (trained using normal commands and HVAC commands).

Classifier Name	Logistic Regression	SVM	Random Forest	Gradient Boosting
Success Rate of Bypassing Defenses	85.3%	89.4%	93.0%	96.3%

all these classifiers, the logistic regression classifier exhibits the best performance with the lowest 90.4% success rate of bypassing defenses. When using the other three classifiers, our success rate of bypassing defenses can reach an average as high as 99.2%. This means that nearly all of the HVAC commands would be recognized as normal commands, which demonstrate the classifiers are not robust enough to defend against our HVAC attacks.

We further test our HVAC commands on a stronger defense classifier that trained directly with HVAC commands and normal commands. The 1,000 recorded HVAC samples are also used as the testing set, and the results are shown in Table 5. Even if the training set involves HVAC commands, due to their acoustic features share a similar pattern with normal commands (Figure 7), it would be hard for the classifiers to distinguish them from normal commands. When using the gradient boosting classifier, we can increase the success rate of bypassing defenses to at most 96.3%, while the logistic regression classifier still outperforms others with the lowest 85.3% success rate. However, this success rate of bypassing defenses is much higher compared with our reproduced results shown in Table 2, which is as low as 0.1%. The convincing results demonstrate our HVAC attacks cannot be defeated even if the ASR system is able to get access to our attack flow and gather a large amount of training data.

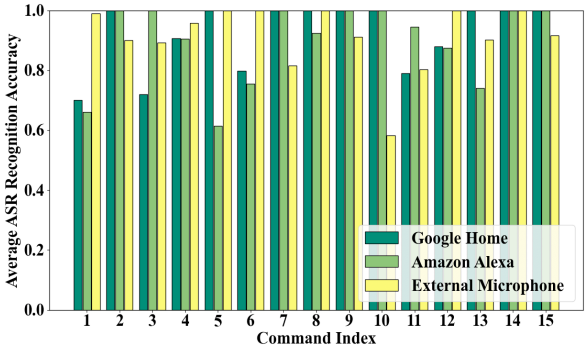


Figure 11: Experimental results of attacking Google Home, Amazon Echo and external microphone under three attack settings.

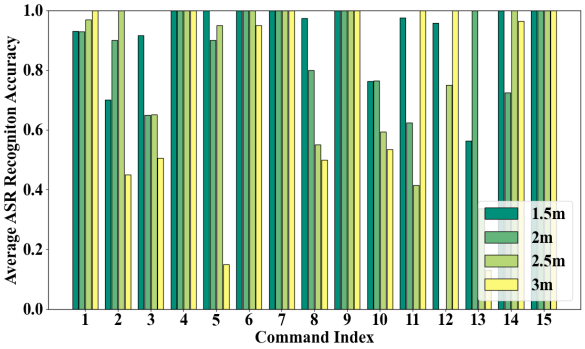


Figure 12: Experimental results for attacking the external microphone with 1.5, 2, 2.5 and 3 meters away.

5.4 Over-the-air Attack Performance

The attack flow shown in Figure 4 ensures the HVAC commands can always be recognized in the digital domain, because an adversary can update the parameters if the sample cannot be recognized after adding the human perception mask. Therefore, we mainly evaluate the performance of our HVAC commands under realistic over-the-air settings.

Attack Performance on Amazon Echo & Google Home. Figure 11 shows the results of attacking Amazon Echo and Google Home smart speaker at setup 1 & 2, respectively. For each speech command, we play the generated HVAC command 20 times, resulting in a total of 300 samples recorded in each setup. We place a sound level meter (i.e., RISEPRO decibel meter) near the speakers to measure the noise level. The ambient noise level is around 36 dBSPL and the commands are played around 70 dBSPL. The recognized transcriptions of these commands could be checked via the mobile apps of these smart speakers online. The average ASR recognition accuracy of these 20 samples are used as the evaluation metric.

We achieve an 89.3% average ASR recognition accuracy for Amazon Echo and 91.9% for Google Home. Among the 15 commands, 7 of them are fully recognized (reach a 100% average ASR recognition accuracy) by Amazon Echo, while 9 out of 15 commands are fully recognized by Google Home. Note although sometimes the transcription is not fully recognized, the smart speakers would also understand the meaning of it and execute the corresponding operation (i.e., command 1 *"What is my schedule today?"* recognized as *"my schedule today"*). The performance results demonstrate our HVAC commands can effectively attack these state-of-the-art commercial smart speakers.

Attack Performance on External Microphone. Figure 11 also shows the results of using an external omnidirectional microphone to record the HVAC commands at a distance of 1 meter from the loudspeaker (setup 3). We feed the recorded samples to the Google speech recognition system API [16] to obtain the recognized transcription. We achieve an overall average ASR recognition accuracy of 91.7%, which is similar to the results obtained from two smart speakers. To further demonstrate the robustness of these commands, we extend the distance between the loudspeaker and the microphone to at most 3 meters, and the results are shown in Figure 12 and Figure 13. Specifically, we achieved an average ASR recognition accuracy of 91.8% at 1.5 meters, 81.9% at 2 meters, 81.4% at 2.5 meters, and 74.6% at 3 meters. As shown in Figure 13, with the increase of the distance between the loudspeaker and the microphone, the average ASR recognition accuracy tends to drop. This is because the sound level of the played commands would decrease at the microphone's side (i.e., drop from 70dBSPL at 1m to 55dBSPL at 3m), meanwhile a larger distortion will be introduced in the prolonged transmission channel. However, even if we increased the distance up to 3 meters, we still achieve a 74.6% overall ASR recognition accuracy which means most of the HVAC commands could be understood, with 7 out of 15 reach 100% average ASR recognition accuracy. These results demonstrate the generalizability of our HVAC commands under various realistic attack scenarios.

6 DISCUSSION

Why Hidden Voice Commands?: In this work, we focus our efforts on advancing the existing hidden voice commands attack, and defeating the existing defense classifier. We find that the bulk of existing research related to hidden voice commands actually focuses on the adversarial example attacks. Although similar, we chose to investigate hidden voice commands for a few significant reasons. First, the work by Carlini *et al.* [6] introduced the first true hidden voice command attack that could defeat ASR while remaining unintelligible to humans and has gained significant attention in academia and media coverage. Additionally, the authors present a modern defense classifier that can defeat their attack as the leading defensive work in this area of research. This encouraged us to explore if and how we could improve the existing hidden voice command attack and its practicality in a real-world setting by defeating the current defense classifier.

Another reason that we chose to focus on hidden voice commands is because in a real-world scenario an adversarial example attack would be very hard to implement. In an adversarial example attack, the attacker insert an audio perturbation over a live issued user command so that it is transcribed to a different command that is chosen by the attacker. Therefore, an attacker would have to remain undetected by the nearby user and issue the correct audio perturbation (specialized for each unique user command) at the same time as the user speaks their live command. So not only will

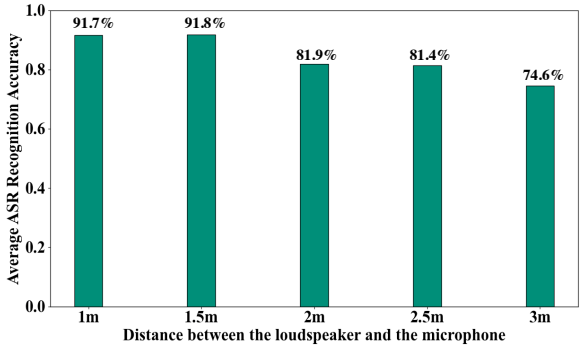


Figure 13: Performance of attacking the external microphone at various distances.

the attacker have to correctly guess what live command the user will speak, but they will also have to initiate the playback of their audio perturbation at the same time as the user starts speaking. On the other hand, in a hidden voice command attack the attacker gets to inject a completely new command to the target system and does not rely on an initial live user command. Since an attacker constructs their hidden voice command prior to attacking the system, they can issue it when the user is not speaking and it should go unnoticed. Further, an attacker could even wait for a more optimum time, when the user is away, to issue the command and reduce the chances of being detected.

Potential Defenses: While our proposed HVAC commands can defeat the existing defense classifier proposed in prior work [6], and even a classifier specially trained on HVAC commands, we recognize that some defense approaches may potentially be used to mitigate our attack. First, *liveness detection* refers to determining whether or not an audio signal originated from a live human speaker or rendered from a speaker device by comparing specific audio features. The work by Zhou *et al.* [39], described in Section 7, demonstrates the potential for liveness detection as a defense strategy. This technique seems robust against all acoustic adversarial command attacks, no matter how imperceptible they are to human listening. As our HVAC commands are still acoustic, and rendered from a speaker device in the attack setting, liveness detection may be a viable defense strategy against our attack.

The defense mechanism proposed by Wang *et al.* [36] also utilizes physical features (i.e., induced vibrations) to distinguish hidden voice commands and normal commands. They determine the unique signatures of each type of command, in the vibration domain, to label any new future commands and identify the hidden voice commands. Again, this technique seems robust against all acoustic adversarial command attacks. An attacker would have to change the device used to playback the adversarial commands and control the induced physical effects from acoustic audio.

Future Work: Recent research has demonstrated the potential for hidden voice command attacks to compromise ASR systems. Many different studies have explored how to craft these commands in unique ways to achieve specific attack goals (i.e., speech recognition, speaker identification, etc.). Conversely, there is very little prominent research on defending against these attacks in a practical setting. This leaves a large area of this research topic in need of further exploration. Future work related to hidden voice commands [6] and our HVAC commands could explore the potential for the above defense techniques to mitigate our attack in controlled experimental settings. Exploring these defense strategies may reveal which of the present techniques is more robust to these hidden voice command attacks.

7 RELATED WORK

Hidden Voice Commands: Most relevant to our research are the recent studies that explore attacks using specialized, *hidden* voice commands. The two main properties of these commands are 1) that they can be recognized by ASR (machine-recognition) and 2) that they are unintelligible to human listeners. The pioneer work on this attack, and the main study that influences our research, is Hidden Voice Commands [6] by Carlini *et al.* In their work, they utilized the reverse feature extraction presented in [35] to construct their commands and extended to a black-box approach. Their work was able to achieve the unintelligibility required for these adversarial commands, and could also beat modern ASR systems.

Other unique works attempt to tackle the "hidden" characteristic of these specialized adversarial examples in new ways. Making an adversarial command more imperceptible (hidden) to human listeners will increase its likelihood of success in an attack. First, Dolphin Attack [38], presented by Zhang *et al.*, exploits the inaudible frequency range above the threshold for human hearing. As a novel approach, the authors modulate their adversarial commands above 20 kHz and find that although the audio signals cannot be detected by the human ear, the ultrasonic frequencies are still capable of relaying speech information to a microphone (for ASR). Similarly, a work by Roy *et al.* [32] exploits the non-linearity of microphone sensors and found a way to design high frequency, inaudible audio signals that can be recorded by standard microphones. The authors present the defensive applications of their crafted signals. Lastly, CommanderSong [37] is a work by Yuan *et al.* that presents a novel technique to hide an adversarial command in the audio of a song. Again, this would allow an attacker to issue an adversarial command in the presence of a victim user without necessarily alerting them to an attack.

Currently, there are few published works that explore the defenses against hidden adversarial commands. Unique to other related works in this area, the Hidden Voice Commands [6] study described previously also evaluated potential defense mechanisms to mitigate hidden voice attacks. Further, they actually presented a defense classifier that can beat their novel hidden voice commands. Inspired by this, we found ourselves asking; *could this defense classifier potentially be beat by even more advanced hidden voice commands?*

A work by Zhou *et al.* [39], recently investigated the vulnerability of VCSs in autonomous cars to the Dolphin Attack [38] and CommanderSong attack [37]. The authors designed a defense mechanism, that utilized *physical measures*, to identify commands spoken by a live human vs. commands played from a speaker device. Specifically, the authors identified the "pop noise" that is produced by human speech when it is near a microphone and used it to distinguish live speech from machine-rendered. In a similar work by Wang *et al.* [36], the authors used learning-based methods and data collected using low-cost MEMS motion sensors (i.e., accelerometer,

gyroscope, etc.) to design a system that can identify hidden voice commands vs. normal commands based on their unique signatures in the vibration domain. So far, the defensive strategies that exploit physical characteristics to distinguish between hidden and normal commands seem to be the most promising for mitigating such attacks.

Research on hidden voice command and adversarial example attacks is continuing to gain popularity and has lead to some published works that provide a Systematization of Knowledge (SoK) on these hidden command attacks and the potential defenses. In a very recent work by Abdullah *et al.* [2], the authors evaluate and systematize a large set of existing research papers on attacks against ASR and speaker identification (SI). They performed experimental tests to assess the transferability of the existing attacks. Recognizing that there has been little published work in the defense space of this research, Abdullah *et al.* discussed *adversarial training* which is a popular defense mechanism developed from related adversarial image research. They conclude that this defense strategy may not be effective in mitigating this attack in the audio domain. The authors also analyzed the current published mechanism for detecting machine vs. human-rendered speech and found that it has true potential as a defense strategy.

Adversarial Examples: Many research works have explored the use of adversarial examples to attack machine learning models. Cocaine Noodles [35], by Vaidya *et al.*, was one of the first published works on generating adversarial examples. In their work, the authors demonstrate a novel (reversed) feature extraction technique that can be used to craft these examples. The authors confirmed that their hidden commands could successfully attack ASR on a smartphone. An interesting work by Carlini *et al.* [7] devised a method to hide their hidden commands in normal audio such that the adversarial example is 99% similar to the human ear as the original audio. This means a malicious command, say "open the garage door", could be hidden in the inconspicuous audio of a TV commercial that would not necessarily alert a victim user that an attack is occurring.

Other works on adversarial commands sought to achieve more specialized goals. In Houdini [9], Cisse *et al.* generated adversarial examples using the loss function of a speech recognition model. Their commands were specialized for combinatorial and non-decomposable speech recognition tasks. Continuing, a work by Iter *et al.* [21] showed that principles of image recognition task models are effective for creating adversarial commands that can defeat speech recognition. An interesting work by Gong *et al.* [14] built an end-to-end scheme to generate adversarial examples by perturbing the audio and modulating it to be unintelligible to humans. The adversarial commands created in this work were designed to target computing paralinguistic applications. Similarly, Alzantot *et al.* [3] designed targeted adversarial attacks against speech classification models. Focusing on short audio clips that classify to one word (i.e., "Yes", "No", "Up", "Down", etc.), the authors forged certain audio features in these clips to get them to classify as a different label. More recently, Abdullah *et al.* [1] generated hidden voice commands for more practical black-box attack settings.

8 CONCLUSION

This paper proposed a more advanced hidden voice attack, HVAC, which can bypass existing learning-based defense classifiers while preserving all the essential characteristics of hidden voice attacks (e.g., unintelligible to humans, recognizable to machines). Existing classifier-based defenses largely rely on the acoustic features extracted from the entire audio of normal and obfuscated samples, whereas only speech parts (i.e., human voice parts) of these samples contain the useful linguistic information for the machine transcription. In this work, we thus proposed a fusion-based method to combine the normal sample and corresponding obfuscated sample as a hybrid command for bypassing these defense classifiers. In addition, several acoustic parameters, including speed, pitch, etc., are further tuned to make the command even harder to be understood by human listeners, yet can still be recognized by machines. Extensive physical over-the-air experiments demonstrated that our proposed HVAC commands can bypass existing defense classifiers while still maintaining a low intelligibility level to humans and a high recognition rate to ASR systems.

ACKNOWLEDGMENTS

We would like to thank our anonymous reviewers for their insightful feedback. This work was supported in part by NSF grants CCF2028876, CCF1909963, CNS1801630, CNS1714807, CNS1526524, CNS1547350 and ARO grant W911NF-18-1-0221.

REFERENCES

[1] Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin RB Butler, and Joseph Wilson. 2019. Practical hidden voice attacks against speech and speaker recognition systems. arXiv preprint arXiv:1904.05734 (2019).
[2] Hadi Abdullah, Kevin M. Warren, Vincent Bindschaedler, Nicolas Papernot, and Patrick Traynor. 2020. SoK: The Faults in our ASRs: An Overview of Attacks against Automatic Speech Recognition and Speaker Identification Systems. ArXiv abs/2007.06622 (2020).
[3] Moustafa Alzantot, Bharathan Balaji, and Mani Srivastava. 2018. Did you hear that? Adversarial Examples Against Automatic Speech Recognition. arXiv:cs.CL/1801.00554
[4] Amazon. 2020. Amazon Echo & Alexa Devices. https://www.amazon.com/smart-home-devices/b?ie=UTF8&node=9818047011.
[5] Apple. 2020. Siri does more than ever. Even before you ask. https://www.apple.com/siri/.
[6] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. 2016. Hidden Voice Commands. In 25th USENIX Security Symposium (USENIX Security 16). USENIX Association, Austin, TX, 513–530. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/carlini
[7] Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 1–7.
[8] Tao Chen, Longfei Shangguan, Zhenjiang Li, and Kyle Jamieson. [n.d.]. Metamorph: Injecting Inaudible Commands into Over-the-air Voice Controlled Systems. ([n. d.]).
[9] Moustapha M Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. 2017. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In Advances in neural information processing systems. 6977–6987.
[10] Statista Research Department. 2020. Number of voice assistants in use worldwide 2019-2023. https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/
[11] Jonathan Driedger and Meinard Müller. 2014. TSM Toolbox: MATLAB Implementations of Time-Scale Modification Algorithms.. In DAFx. Citeseer, 249–256.
[12] EDIFIER. 2020. Amp up your audio R980T 2.0 Active Speaker System. https://www.edifier.com/us/en/speakers/r980t-2.0-powered-bookshelf-speakers.
[13] Theodoros Giannakopoulos. 2015. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. PloS one 10, 12 (2015).
[14] Yuan Gong and Christian Poellabauer. 2017. Crafting Adversarial Examples For Speech Paralinguistics Applications. arXiv:cs.LG/1711.03280

[15] Google. 2020. Google smart speaker and displays. https://store.google.com/us/magazine/compare_nest_speakers_displays?srp=/us/product/google_home.
[16] Google. 2020. Google Speech-to-Text. https://cloud.google.com/speech-to-text.
[17] Google. 2020. Set up payments for Google Assistant on Google Nest or Google Home speaker or display. https://support.google.com/googlenest/answer/7276665?hl=en.
[18] Theodore D Hanley and Mack D Steer. 1949. Effect of level of distracting noise upon speaking rate, duration and intensity. Journal of Speech and Hearing Disorders 14, 4 (1949), 363–368.
[19] Aki Harma. 2003. Automatic identification of bird species based on sinusoidal modeling of syllables. In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)., Vol. 5. IEEE, V–545.
[20] Md Rashidul Hasan, Mustafa Jamil, MGRMS Rahman, et al. 2004. Speaker identification using mel frequency cepstral coefficients. variations 1, 4 (2004).
[21] Dan Iter, Jade Huang, and Mike Jermann. 2017. Generating adversarial examples for speech recognition. Stanford Technical Report (2017).
[22] Matěj Korvas, Ondřej Plátek, Ondřej Dušek, Lukáš Žilka, and Filip Jurcícek. 2014. Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 4423–4428.
[23] Matěj Korvas, Ondřej Plátek, Ondřej Dušek, Lukáš Žilka, and Filip Jurčíček. 2014. Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license. In Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2014). To Appear.
[24] Paige Leskin. 2018. Here's how to use Duplex, Google's crazy new service that impersonates a human voice to make appointments on your behalf. https://www.businessinsider.com/google-appointment-booking-reservation-tool-duplex-pixel-phones-how-to-2018-11
[25] Matt May. 2005. Inaccessibility of CAPTCHA: Alternatives to visual Turing Tests on the Web. web page. URL: http://www. w3. org/TR/turingtest (2005).
[26] K. Naithani, V. M. Thakkar, and A. Semwal. 2018. English Language Speech Recognition Using MFCC and HMM. In 2018 International Conference on Research in Intelligent and Computing in Engineering (RICE). 1–7. https://doi.org/10.1109/RICE.2018.8509046
[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12 (2011), 2825–2830.
[28] John F Pitrelli, Raimo Bakis, Ellen M Eide, Raul Fernandez, Wael Hamza, and Michael A Picheny. 2006. The IBM expressive text-to-speech synthesis system for American English. IEEE Transactions on Audio, Speech, and Language Processing 14, 4 (2006), 1099–1108.
[29] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. 2019. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In International Conference on Machine Learning. 5231–5240.
[30] Douglas A. Reynolds. 1995. Speaker identification and verification using Gaussian mixture speaker models. Speech Communication 17, 1 (1995), 91 – 108. https://doi.org/10.1016/0167-6393(95)00009-D
[31] Phil Rose. 2002. Forensic speaker identification. cRc Press.
[32] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. 2017. BackDoor: Making Microphones Hear Inaudible Sounds. In Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '17). Association for Computing Machinery, New York, NY, USA, 2–14. https://doi.org/10.1145/3081333.3081366
[33] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. 2018. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. arXiv preprint arXiv:1808.05665 (2018).
[34] Telsa. 2020. Telsa Voice Commands. https://www.tesla.com/support/voice-commands.
[35] Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. 2015. Cocaine Noodles: Exploiting the Gap between Human and Machine Speech Recognition. In 9th USENIX Workshop on Offensive Technologies (WOOT 15). USENIX Association, Washington, D.C. https://www.usenix.org/conference/woot15/workshop-program/presentation/vaidya
[36] Chen Wang, S Abhishek Anand, Jian Liu, Payton Walker, Yingying Chen, and Nitesh Saxena. 2019. Defeating Hidden Audio Channel Attacks on Voice Assistants via Audio-Induced Surface Vibrations. In Proceedings of the 35th Annual Computer Security Applications Conference (ACSAC '19). Association for Computing Machinery, New York, NY, USA, 42–56. https://doi.org/10.1145/3359789.3359830
[37] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, Xiaofeng Wang, and Carl A. Gunter. 2018. CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition. In USENIX Security Symposium.
[38] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. Dolphinattack: Inaudible voice commands. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 103–117.
[39] Man Zhou, Zhan Qin, Xiu Lin, Shengshan Hu, Qian Wang, and Kaili Ren. 2019. Hidden Voice Commands: Attacks and Defenses on the VCS of Autonomous Driving Cars. IEEE Wireless Communications PP (04 2019), 1–6. https://doi.org/10.1109/MWC.2019.1800477

A APPENDIX

A.1 Command List

Table 6: HVAC Command List

Command Index	Command
1	What is my schedule today?
2	Take a picture.
3	Turn off all lights.
4	Turn on the airplane mode.
5	Ok google.
6	Call 911.
7	Open the door.
8	I cannot find my book.
9	What is my current location?
10	Turn off all alarms.
11	Play some music.
12	Turn the volume up.
13	What is the weather today?
14	Set a timer for 5 minutes.
15	Wake me up at 8 a.m.