BiasHacker: Voice Command Disruption by Exploiting Speaker Biases in Automatic Speech Recognition

Payton Walker
prw0007@tamu.edu
Texas A&M University
College Station, Texas, USA

Nathan McClaran
nmcclaran@tamu.edu
Texas A&M University
College Station, Texas, USA

Zihao Zheng
zihaozheng@tamu.edu
Texas A&M University
College Station, Texas, USA

Nitesh Saxena
nsaxena@tamu.edu
Texas A&M University
College Station, Texas, USA

Guofei Gu
guofei.gu@tamu.edu
Texas A&M University
College Station, Texas, USA

ABSTRACT

Modern speech recognition systems that are widely deployed today still suffer from known gender and racial biases. In this work, we demonstrate the potential to exploit the existing biases in these systems to achieve a new attack goal. We consider the potential for command disruption by an attacker that can be conducted in a manner that allows for access and control of a victim's voice assistant device. We present a novel attack, BiasHacker, which crafts specialized chatter noise to exploit racial and gender biases in speech recognition systems for the purposes of command disruption. Our experimental results confirm both racial and gender bias that is still present in the speech recognition systems of two modern smart speaker devices. We also evaluated the effectiveness of three types of chatter noise (American English (AE)-Male, Nigerian-Female, Korean-Female) for disruption and demonstrate that the AE-Male chatter is consistently more successful. Comparing the average success rate of each chatter type, in scenarios where disruption was achieved, we find that when targeting the Google Home mini smart speaker, the AE-Male chatter noise increases average disruption success compared to the Nigerian-Female and Korean-Female chatter noises by 112% and 121%, respectively. Also, when targeting the Amazon Echo Dot 2 the AE-Male chatter noise increases average disruption success compared to the Nigerian-Female and Korean-Female chatter noises by 42% and 69%, respectively.

CCS CONCEPTS

• **Computing methodologies** → **Speech recognition**; • **Security and privacy** → **Denial-of-service attacks**.

KEYWORDS

speech recognition, bias, command disruption, voice assistant

WiSec '22, May 16–19, 2022, San Antonio, TX, USA

ACM ISBN 978-1-4503-9216-7/22/05…$15.00
https://doi.org/10.1145/3507657.3528558

ACM Reference Format:
Payton Walker, Nathan McClaran, Zihao Zheng, Nitesh Saxena, and Guofei Gu. 2022. BiasHacker: Voice Command Disruption by Exploiting Speaker Biases in Automatic Speech Recognition. In *Proceedings of the 15th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec '22), May 16–19, 2022, San Antonio, TX, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3507657.3528558

1 INTRODUCTION

Racial and gender biases in recognition systems are a known issue even among current state-of-the-art technology and these biases persist in speech recognition systems [11, 14, 18]. There are a number of reasons in the machine learning models that can cause these outcomes and lead to low accuracies among certain speaker groups. There can be feedback problems from an original biased model that led to biases in a future model, the learning algorithm can be biased because it is optimized for overall error rate, the ground truth may not be labeled accurately for marginalized groups, the marginalized group could be poorly represented in the data, or there could be labeling issues. Until these are resolved, the harm caused by these biases will continue and possibly lead to new vulnerabilities that can affect a user. This paper explores one such vulnerability that allows an attacker to exploit biases in speech recognition systems for a specialized attack.

Voice assistant technology, such as smart speakers, is becoming more commonplace in the household and new threats to the users are emerging. The typical attack methodology for controlling a smart home environment, up to this point, has been focused on command injection attacks [2, 15, 19, 20]. Since smart speakers do not perform speaker verification by default for all commands, virtually anyone could issue a command to a smart speaker as long as they can gain access. However, these injection attacks have significant challenges such as sophisticated processes for creating the obfuscated command and requiring that the command be unrecognizable to human listeners while remaining understandable to machines, often making them impractical in live attack situations [1]. Attacks that seek to disrupt a user's command and keep it from being recognized are less common. Most of the work that looks to block a user's command has an overall Denial of Service (DoS) goal in which all functions of the virtual assistant are kept from the user. In this work, we acknowledge the threat of disrupting user commands as another form of DoS and realize the scenarios in

which this poses a significant risk to the user. In particular, blocking a user's command to lock a door can give an attacker access to a home, or blocking the "hangup" command on an active call between smart speakers can leave the call connection open for an attacker to eavesdrop.

Aside from the typical approach for blocking a user command that utilizes some type of jamming signal, and in consideration of the known racial and gender biases, we consider the possibility of feeding specialized chatter noise (speech + music + noise) to the speech recognition system that would be prioritized and recognized in place of the legitimate user speech. For example, if noise (possibly played through a TV or radio) containing speech from an American English (AE) speaking male is played over a legitimate voice command given by a female speaker with a non-AE accent, the smart speaker may recognize the male speech content and attempt to transcribe it without acknowledging the legitimate command from the user. We chose a chatter type noise because we found that dynamic noises such as speech and music are not filtered by noise cancellation, while static noises are easily filtered. The music and speech combination in the chatter noise work to make it inconspicuous to a user. And since the noise sample will be short, in order to mask just a particular word or command, it is likely the noise will go unnoticed or be considered benign. In this light, we present the BiasHacker attack which exploits known speaker recognition biases in order to disrupt a user's command (shown in Figure 1). To our knowledge, we are the first academic study to consider this broad attack vector for targeting speech recognition systems. Notably, this methodology could apply to many domains.

Main Contributions and Results: We summarize our key contributions and results below:

(1). Bias-based Disruption Noise: We generated specialized noise that is designed to exploit known biases in speech recognition in order to improve attack success. We evaluate three different types of chatter noise that uses speech content from three different types of speakers (AE-Male, Nigerian-Female, Korean-Female). For the purposes of the attack, we suspect the AE-Male chatter noise will perform best at disrupting user commands because it contains speech that is known to be better represented and therefore preferred by speech recognition systems.

(2). Black-Box Analysis: We performed black-box experiments using two live implementations of voice assistant technology including a Google Home mini and an Amazon Echo Dot 2. The experiments were conducted to evaluate the potential of the different types of disruptive chatter noise. We confirm the presence of racial bias where non-AE accented speech is more easily disrupted compared to American English speech (found in the Google device where only non-AE accented speech samples were successfully disrupted). We also confirm the presence of gender bias where female speech is more easily disrupted compared to male speech (found in the Amazon device where only female speech samples were successfully disrupted). Lastly, we demonstrate that a chatter noise containing American English male speech is more effective at command disruption than using speech from a Nigerian female or Korean female which are less represented in speech recognition systems (e.g., female speaker with a non-AE accent).

2 THREAT MODEL

In our threat model, the attacker intends to block a user's command from being accepted by their voice assistant (VA) device. The attacker targets users from under-represented groups because against these users, disruption attacks can be exploited more effectively. Using external noise injection, the attacker plays the disruption noise in the same environment as the victim. This can be accomplished by multiple channels including, playing the noise from a loudspeaker outside the space where the victim is located. Consider a disruption noise that is disguised as a lawn-mower sound and played outside a window. In this scenario the noise would be inconspicuous to a user, but could remain effective at hindering the VA device from accepting the user's command. Another possibility is that the attacker can play their disruption noise through a speaker device inside the room such as a television (noise disguised by a commercial) or radio (noise disguised as music).

While most existing smart speaker attacks are aimed at injecting a command [2, 15, 19, 20], blocking a command is a more practical threat because injection can be quite difficult. These attacks typically involve some form of obfuscated voice command that requires sophisticated techniques to generate. Additionally, it has a requirement to be unrecognizable by a human listener so that it may go undetected. With these hurdles, tackling a command injection attack may not be feasible. We can identify multiple scenarios where blocking a command would also pose a significant threat. This is especially the case since smart home environments are becoming more popular where multiple smart devices are connected and controllable from a central VA device. Consider a user who is about to go to bed and issues a command to lock their doors or turn on security cameras. If these commands are blocked, an attacker may be able to gain access through the unlocked doors and go undetected by the cameras. One of the more threatening scenarios for command blockage, which we use as a representative example in this work, is an active call between devices that the user attempts to hangup. If the attacker blocks the hangup command without the user's knowledge, the connection between devices will remain open and active. In this scenario the attacker can compromise that connection and use it to issue any command to the victim's device, or use the open connection to listen in on the victim.

3 METHODOLOGY

3.1 Experimental Setup

Our experiments were performed in a quiet room with no other noises in the environment that could affect our results. As our experiments are designed to evaluate command disruption using unique samples of noise, we maintained a consistent setup with two different audio sources (one for the normal command audio and one for the noise audio) including a portable loudspeaker and a separate loudspeaker system. In all experiments the loudspeakers were connected to two different cell phones that stored the audio files, and we placed each on opposite sides of the smart speaker facing towards it. The loudspeakers were located 0.5 meters from the smart speaker and a digital sound level meter was used to ensure the normal command and noise were played at the correct volume for each experiment (measured at the location of the smart speaker). A lab member manually activated the normal command audio and

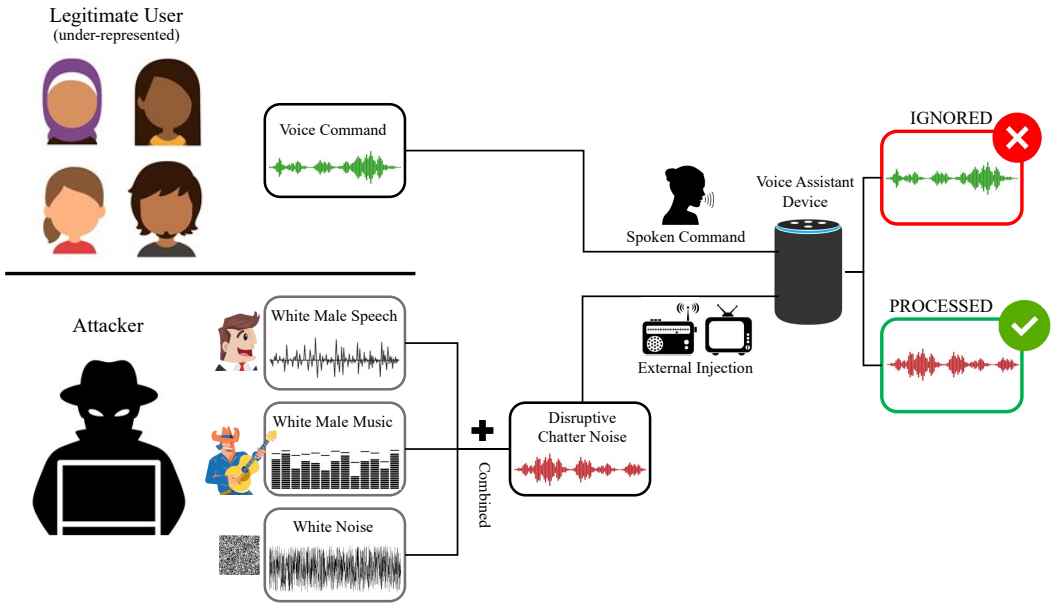

Figure 1: Flow diagram depicting the steps in the BiasHacker attack. The attacker generates a disruptive chatter noise combining raw speech and music from a preferred speaker type (i.e., American English male) with Gaussian White Noise. The chatter noise is then played when the legitimate user speaks their command and the smart speaker attempts to interpret the chatter noise and ignore the legitimate command. If the legitimate user is from a community that is biased in a speech recognition model, the preferred speaker in the chatter noise will be prioritized and the actual user command will be blocked.

noise sample so that the noise was played over the command portion only. Using indicators from the smart speaker and checking online voice command history logs, we recorded the number of times the command was disrupted for each scenario.

3.2 Experimental Parameters

Sound Pressure Level (SPL): We test the noise audio at different loudness levels to understand when it is possible to disrupt smart speaker commands. In our experiments the normal commands are played at 70 dB which is above the loudness level for normal human conversation (40-60 dB), but is representative of the raised, presentation style voice that most people use when addressing a smart speaker device. For the disruptive noise we test loudness levels up to the same level that the normal command is played. Specifically, we test the disruptive noise at 50, 60 and 70 dB. Considering a practical attack scenario, if the disruptive noise is louder than the normal command spoken by the user, it would almost certainly be detected and would raise alarm to the user that their command may not have been understood.

Victim Speaker Type: We consider multiple speaker types in our representation of potential victim speakers. We include both female and male speech samples encompassing five different accents for English speech, elaborated in Section 3.4. One of the speaker accents is American English, representing the main speaker type that automatic speech recognition models perform the best for. The other five speaker types encompass non-AE language accents that affect the pronunciation of English words and can pose a challenge for speech recognition systems.

Disruptive Noise Type: We generate three different types of disruptive chatter noise, elaborated in Section 3.5, to observe whether known speech recognition bias can be exploited for the purposes of the BiasHacker attack. One of the noise types represents the gender and accent of speech known to be preferred by smart speakers (AE, male), while the other two types represent speakers that may be less preferred (non-AE accent, female). Since American English male data is often well represented, it is reasonable to assume that a masking noise which contains this type of speech will perform better at disrupting a user command (e.g., speech contained in the chatter noise is preferred over the user's own speech).

3.3 Equipment

The two loudspeaker devices used in our experiments were a Sony SRS-XB2 Bluetooth portable speaker (for the normal command audio) and a Logitech Z323 speaker system (for the noise audio). We used a Rolls SLM305 digital sound level meter to ensure the normal command and noise audio were played at the correct dB level in each experiment. For our victim VA we used the Google Home mini (wake word: "Hey Google") and Amazon Echo Dot 2 (wake word: "Alexa") smart speakers.

3.4 Speech Dataset

For our normal command audio samples of the "hangup" command we used a Python script and the Google Text-to-Speech API to generate audio samples in English for five different accents. For plain American English we generated three female and three male speaker samples (en-US-Wavenet-(A-F)). For the remaining four accents we generated two female and two male speaker samples. The non-AE accents used included Australian (en-AU-Wavenet-(A-D)), German (de-DE-Wavenet-(A-D)), British (en-GB-Wavenet-(A-D)), and Korean (ko-KR-Wavenet-(A-D)). Therefore, for each smart

Table 1: Summary of command disruption results for each speaker accent and gender, chatter noise type, and both smart speakers. "✓": Successful command disruption, "✗": No command disruption.

Accent	Speaker Gender	Amazon Echo Dot 2			Google Home mini		
		Chatter Type			*Chatter Type*		
		AE-Male	Nigerian-Female	Korean-Female	AE-Male	Nigerian-Female	Korean-Female
Australian	F	✓	✓	✓	✓	✓	✓
	M	✗	✗	✗	✓	✗	✗
German	F	✓	✗	✗	✓	✓	✓
	M	✗	✗	✗	✓	✓	✓
British	F	✓	✗	✗	✓	✗	✗
	M	✗	✗	✗	✓	✗	✗
Korean	F	✓	✗	✗	✓	✓	✓
	M	✗	✗	✗	✓	✓	✓
American English (AE)	F	✓	✓	✓	✗	✗	✗
	M	✗	✗	✗	✗	✗	✗

speaker, we used a total set of 22 speaker samples. Before beginning our official experiments, we confirmed that each of the normal command audio samples were recognized 100% of the time by each of the smart speakers, in the absence of noise.

3.5 Bias-based Disruption Noise Generation

We created three different versions of chatter noise (termed *AE-Male*, *Nigerian-Female*, and *Korean-Female*) that are designed to reveal the potential for exploiting speech recognition bias in our attack. For each chatter noise we combined a sample of plain speech (from TedTalks), music (from Youtube), and Gaussian White Noise. These are the basic building blocks of a chatter noise. The plain speech samples were all spoken in English with the Nigerian female speaker having a Nigerian accent and the Korean female speaker having a Korean accent. The music samples selected for each chatter noise also included speech of the appropriate dialect/language. For the AE-Male chatter noise we used a popular Country song, for the Nigerian-Female chatter noise we used a pop song from a Nigerian female pop singer, and for the Korean-Female chatter noise we used a ballad sung by a Korean girl (in the Korean language). From each of these chatter noises, we isolated three one second clips for a total of nine different chatter noise samples.

4 RESULTS

In this section we present our results for the BiasHacker attack. We conducted a large set of experiments that test the three versions of the bias-based chatter noise that we created (AE-Male, Nigerian-Female, Korean-Female). Using the normal command audio sample from each TTS speaker (described in Section 3.4), we attempted 10 command disruption attempts for each version of disruption noise (nine total). We recorded the percentage of attempts that the normal command was successfully disrupted (e.g., incorrectly recognized or not recognized). We ran experiments to test the disruptive chatter noise samples at the 50, 60, and 70 dB SPL levels. We found that at 50 dB, none of the chatter noises were able to disrupt any of the normal command audio samples. Conversely, at the 70 dB SPL level, we found that all samples of the chatter noise were 100% successful at disrupting all of the normal command audio samples (e.g., noise was too loud to observe potential bias exploitation). Therefore, we present our findings for the 60 dB SPL level because there are instances of both success and failure for command disruption. Table 1 shows the summarized results for attack success/failure for each speaker type, chatter version, and smart speaker. We refer to our project website (https://sites.google.com/view/bias-hacker) for a full set of our results including accuracy tables, success summary, and descriptions of the preliminary experiments and analysis that were conducted.

4.1 Evidence of Racial Bias

The presence of racial bias (e.g., bias against non-AE accents) is very apparent in the results for the Google Home mini. We observed up to 100% command disruption success against the Australian, German, and Korean speakers. And we observed up to 80% command disruption success against the British speakers. On the other hand, there were no successful disruption attempts against the American English speakers (male or female) for any of the chatter noise types. These results suggest that racial bias persists in the speech recognition system used by Google. We did not observe the same racial bias in the Amazon Echo Dot 2 results. In fact, we were surprised to find that the American English speaker samples were the easiest to disrupt compared to all the others.

4.2 Evidence of Gender Bias

In the results for the Amazon Echo Dot 2, we can see evidence of gender bias in the speech recognition system used by Amazon. Specifically, there was some disruption success against female speakers of all accent types. We observed up to 60% disruption success against female speakers with non-AE accents, and up to 100% disruption success against American English female speakers. Interestingly, we did not observe any disruption success against any of the male speakers with any of the chatter noise samples in the attacks against the Amazon Echo Dot 2. Comparatively, we did not observe the same gender bias in the Google Home mini results, where both male and female samples were successfully disrupted.

4.3 Evidence of Bias Exploitation via Noise

Once we averaged the command disruption rates for each type of chatter noise, it becomes clearer that there is a real potential for

exploiting the known biases for the purposes of an attack. From our results, we find a consistent trend concerning the effectiveness of each chatter noise type to disrupt user commands. We find that between all three types of chatter noise, the AE-Male chatter is consistently the best performing for command disruption compared to the Nigerian-Female and Korean-Female chatter. This observation holds true in the results for both the Google Home mini and Amazon Echo Dot 2. Between the Nigerian-Female and Korean-Female chatter there is no clear pattern that suggests one type works better than the other. However, the success rates for both types are always less than the AE-Male chatter (in the instances where disruption success occurs). Based on the previous observations described above, it seems that the AE-Male chatter noise exploits the *racial* bias in the Google Home mini, where the other chatter noise samples use non-AE accented content. It also exploits the *gender* bias in the Amazon Echo Dot 2, where the other chatter noise samples use female content. In both cases, the AE-Male chatter is revealed to be the most effective.

If we consider the average success rates in scenarios where disruption was achieved, we can see how much more effective the AE-Male chatter noise is for the attack. Against the Google Home mini, the average success rate of the AE-Male chatter is 53%. This is a 112% increase from the average success rate for the Nigerian-Female chatter (25%), and a 121% increase from the average success rate for Korean-Female chatter (24%). Against the Amazon Echo Dot 2, the average success rate of the AE-Male chatter is 27%. This is a 42% increase from the average success rate for the Nigerian-Female chatter (19%), and a 69% increase from the average success rate for the Korean-Female chatter (16%). This strongly suggests that the biases existing in modern speech recognition systems (e.g., preference to American English male speech) can be exploited for command disruption attacks.

5 FURTHER INSIGHTS

We observed clear indications of both racial and gender biases in the smart speakers that were used. We found that racial bias exists in the speech recognition system used by Google, and gender bias exists in the speech recognition system used by Amazon. We also observed different behaviors from the smart speaker devices that may lend to the presence of these biases. First, the Amazon device seemed more strict about accepting user commands. If the audio was perturbed in any way from the expected clear command, the Amazon device would ignore it all together. The device would give no response and would not illicit the user to try again. This suggests that strong, clear commands in distinguishable tones are preferred. This could be why male speech, which contains lower frequencies and is less obscured by noise injection, are easier to recognize.

Next, compared to the Amazon device, the Google device seemed to put a greater emphasis on accepting a command (even if it is wrong) compared to not accepting one at all. We observed significantly more instances of mis-recognition (identifying the wrong command) with the Google Home mini which is still considered a command disruption. For example, for the Korean male speakers, the "Hangup" command was often mis-recognized as "King Kong" or "Ping Golf" when the disruption noise was added. This resulted in a response about the King Kong movie, or nearby location suggestions for playing golf. And for the German female speakers, the "Hangup" command was often mis-recognized as "Hangover" or "Array" resulting in a Wikipedia search and definitions for the words. The Google device seems more apt to identify the speech and maintain a dialogue with the user which could explain why foreign accented speech is easier to disrupt. While the clarity of the English is already lessened, adding in additional noise may cause the device to more quickly mis-recognize the speech. So instead of diligently processing the recorded audio, the device will simply accept the first thing that it can recognize and then respond accordingly.

While the BiasHacker attack reveals a significant vulnerability in the exploitation of speech recognition bias, there are a few known limitations to this attack. First, this attack will certainly be less effective, if at all, against speakers that are well represented in the training of the speech recognition systems (i.e., American English speaking men). Since their speech is already the most preferred in these recognition systems, the approach to introduce speech that is more favored does not apply. Another limitation to the attack is the requirement of external noise sources to be near the victim's voice assistant. Without a medium to play the disruptive noise, the attack cannot be executed. Finally, certain scenarios would not be conducive to injecting a disruptive noise without detection. A diligent user may wait for confirmation that their command was accepted and could become immediately aware that their command was disrupted if they pay attention.

In order to defend against the BiasHacker attack, we need a solution to mitigate the effects of the inherent speaker biases in speech recognition systems. Specifically, a solution is needed that can ensure the prioritization of the legitimate user's speech. One solution to this would be a new form of speaker profile setting on voice controllable devices that not only recognizes the trained user's voice profile, but will deny voice commands from any other speaker. Another option could be an external device placed on the voice assistant microphone that synthesizes the users voice in real time to a known speaker type that is preferred (e.g., American English male). Doing this would also exploit the speaker biases, but this time it would be to benefit the user. Lastly, improvements in speaker recognition and source localization on the devices can be used to thwart this attack. If the device can recognize that the legitimate command and the noise are coming from different locations, it will become better at filtering all parts of the disruptive noise and accept the actual voice command.

6 RELATED WORK

Command Injection: Previous work on disrupting smart speaker and voice assistant commands have focused on command injection attacks designed to trick the voice assistant into accepting malicious commands. Various approaches have been taken, including audible attacks using hidden voice commands [2], and inaudible attacks using ultrasonic frequencies with both the air [20] and solid table surface [19] as the transfer medium. Another work presented Light Commands[15], a laser-based attack that encoded commands onto a beam of light. These attacks focused on controlling the voice assistant devices by injecting their own commands and did not consider the potential to block user commands to achieve an attack

goal. In fact, a recent work by Abdullah et al. [1] revealed that much of these works are impractical in real-world setting because they lack tranferability to multiple systems and they had not been tested in over-the-air settings. Changing what a user says on the fly or adding new speech that the user did not say, without them noticing, would be extremely challenging in a live setting. Our work however uses a short clip of noise to disrupt parts of a user command and there are many applicable scenarios where this attack would be practical and could go undetected.

Smart Speaker Jamming/DoS: Various forms of jamming have been considered for use with smart speakers and voice assistants, for both malicious and defensive purposes. In a work by Chandrasekaran et al. [3], the authors presented a privacy preserving technique that used ultrasound jamming to prevent smart speakers from eavesdropping. Similarly, Chang et al. [4] designed a reactive jamming device that could disrupt speaker recognition of a wake word. However, these defenses prevent any interaction with the target system rendering further exploitation useless. Jamming techniques have also been used to enact a DoS attack on the Siri and Alexa voice assistants [13]. These attacks targeted the wake word on devices, and did not explore disruption of critical commands. These studies demonstrated that total jamming may be effective as a brute defensive approach, but are not beneficial from an attack perspective because of the limited control.

Machine Learning Bias: Many previous research works have identified instances of bias that exist in different speech-related systems [5, 8–10, 12]. Tatman and Kasten [16] identified dialect and racial biases in Youtube's automatic caption system that had statistically different word error rate (WER) compared to American English speakers who had the lowest average WER. Garnerin et al. [6] examined four major speech corpora in French broadcasting that were widely used by the speech community to train automatic speech recognition systems. They identified gender bias based on a lack of female representation in TV and radio who then became under-represented in the data. Another work by these authors found a lack of gender information in speech resources from the Open Speech and Language Resource platform which impacted transparency and fairness [7]. They realized that achieving gender balance in the data requires other speech corpus characteristics. In a work by Vanmassenhove et al. [17], the authors found that for certain non-English languages encoding gender information in the data improved the performance of neural machine translation.

7 CONCLUSION

In this work we presented the BiasHacker attack that utilizes specially crafted chatter noise and exploits existing speech recognition biases to disrupt a user's command. This attack can be launched against any voice controllable device to disrupt critical commands. We evaluated the effectiveness of three different types of disruptive chatter noise against an array of different speaker types (male/female, American English/non-AE accented). Our results first confirm the presence of racial bias in the Google Home mini and gender bias in the Amazon Echo Dot 2. We find that AE-Male chatter is more effective, compared to the other chatter noises, for the Google Home mini and Amazon Echo Dot 2 by approximately 100% and 50%, respectively. Future work can look at more sophisticated noises and potentially expand the capabilities of the attack by embedding commands to be injected.

ACKNOWLEDGMENTS

This research is partially supported by the National Science Foundation (NSF) under the grants: CNS-1714807, CNS-2030501, CNS-2139358, and 1816497.

REFERENCES

[1] Hadi Abdullah, Kevin Warren, Vincent Bindschaedler, Nicolas Papernot, and Patrick Traynor. 2021. SoK: The Faults in our ASRs: An Overview of Attacks against Automatic Speech Recognition and Speaker Identification Systems. *2021 IEEE Symposium on Security and Privacy (SP)*, 730–747.

[2] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Michael E. Sherr, Clay Shields, David A. Wagner, and Wenchao Zhou. 2016. Hidden Voice Commands. In *USENIX Security Symposium*.

[3] Varun Chandrasekaran, Kassem Fawaz, Bilge Mutlu, and Suman Banerjee. 2018. Characterizing privacy perceptions of voice assistants: A technology probe study. *arXiv preprint arXiv:1812.00263* (2018).

[4] Peng Cheng, Ibrahim Ethem Bagci, Jeff Yan, and Utz Roedig. 2018. Towards reactive acoustic jamming for personal voice assistants. In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*. 12–17.

[5] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying Bias in Automatic Speech Recognition. *ArXiv* abs/2103.15122 (2021).

[6] Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2019. Gender Representation in French Broadcast Corpora and Its Impact on ASR Performance. *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery* (2019).

[7] Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2020. Gender Representation in Open Source Speech Resources. In *LREC*.

[8] Jan Gorisch, Michael Gref, and Thomas C. Schmidt. 2020. Using Automatic Speech Recognition in Spoken Corpus Curation. In *LREC*.

[9] Chunxi Liu, Michael Picheny, Leda Sari, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. 2021. Towards Measuring Fairness in Speech Recognition: Casual Conversations Dataset Transcriptions. *ArXiv* abs/2111.09983 (2021).

[10] Joshua L. Martin. 2021. Spoken Corpora Data, Automatic Speech Recognition, and Bias Against African American Language: The case of Habitual 'Be'. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021).

[11] Cade Metz. 2020. *There Is a Racial Divide in Speech-Recognition Systems, Researchers Say.* https://www.nytimes.com/2020/03/23/technology/speech-recognition-bias-apple-amazon-google.html

[12] Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell. 2020. Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications. In *LREC*.

[13] Taekkyung Oh, William Aiken, and Hyoungshick Kim. 2018. Hey Siri – Are You There?: Jamming of Voice Commands Using the Resonance Effect (Work-in-Progress). *2018 International Conference on Software Security and Assurance (ICSSA)* (2018), 73–76.

[14] Mishaela Robison. 2020. *Voice assistants have a gender bias problem. What can we do about it?* https://www.brookings.edu/blog/techtank/2020/12/09/voice-assistants-have-a-gender-bias-problem-what-can-we-do-about-it/

[15] Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, and Kevin Fu. 2020. Light Commands: Laser-Based Audio Injection Attacks on Voice-Controllable Systems. In *USENIX Security Symposium*.

[16] Rachael Tatman and Conner Kasten. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Proc. Interspeech 2017*. 934–938. https://doi.org/10.21437/Interspeech.2017-1746

[17] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting Gender Right in Neural Machine Translation. *ArXiv* abs/1909.05088 (2018).

[18] Kyle Wiggers. 2021. *Study finds that even the best speech recognition systems exhibit bias.* https://venturebeat.com/2021/04/01/study-finds-that-even-the-best-speech-recognition-systems-exhibit-bias/

[19] Qiben Yan, Kehai Liu, Qin Zhou, Hanqing Guo, and Ning Zhang. 2020. SurfingAttack: Interactive Hidden Attack on Voice Assistants Using Ultrasonic Guided Waves. In *NDSS*.

[20] 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 (CCS '17)*. Association for Computing Machinery, New York, NY, USA, 103–117. https://doi.org/10.1145/3133956.3134052