THE UNIVERSITY OF WESTERN AUSTRALIA

PROJECT PROPOSAL

Tony Liu 21492926

APRIL 12, 2019

1. Introduction

Voice assistants are software agents which operate on smartphones or purpose-built devices such as smart TVs and speakers which listens for, processes and interprets human speech into commands [1]. Voice assistants are constantly listening for activation words to wake up and process human speech, such as “Siri” for Apple’s Siri. In 2018, U.S. voice assistant owners rose by 40% to a total of 66.4 million users, with Amazon Echo occupying 61% of the market [2]. A study from Juniper Research [3] estimates that there were 2.5 billion voice assistants in use at the end of 2018, with this figure tripling to 8 billion by 2023. The concept of voice assistants was popularised by the 1968 film Space Odyssey, which included a sentient computer which became a pop-culture icon, the HAL9000 [4]. This concept was made a reality, when the app ‘Siri’ was developed in 2010 and integrated by Apple into the iPhone 4s, which was released on October 4, 2011 [5]. However, with all the convenience and functionality brought about by voice assistants, the attraction of its exploitation is inevitable.

A 2015 study [6] demonstrated the gaps between human and machine speech recognition. Vaidya [6] states that machines mimic the way a human processes language, converting analogue signals to digital representations and using artificial intelligence to organize those representations into phonemes. This is exploitable, as obfuscated audio can be interpreted by machines as speech while being undetectable by a human, due to how voice assistants convert signals to text. These attacks can be split into two broad categories: black box or white box tests. A black box test is a type of test without any knowledge of the internal logic and structure of the target system, aiming to understand how a system functions. White box tests are conducted with some knowledge of the system’s source code and functionality, with the goal of understanding the internal logic and structure of the code [7].

Some proposed protection methods include requiring a password being entered to authenticate users, however, this can be counter-intuitive, as the purpose of voice assistant integration is convenience. In many scenarios, it can be impractical to enter a password to authenticate a user. Rather than focusing on manual authentication, passive defence measure can be integrated. Behaviour specification is a potential defence method which determines whether or not a command is normal or malicious.

My intended contribution towards this field of research focusses on the protection of voice assistants against adversarial attacks. This will be achieved by establishing a framework for behavioural specification technology integrated into voice assistants.

2. Literature Review

2.1 Speech Recognition

Voice assistants use a combination of different software to process human speech into executable commands. These can be separated into three core actions: speech-to-text, text-to-intent and intent-to-action [8]. Modern voice assistants utilise deep neural networks (DNN) in combination with speech recognition approaches such as Hidden Markov Models [9].

2.1.1 Speech-to-Text

Speech-to-text conversion is accomplished through the use of automatic speech recognition (ASR) powered by DNNs [10]. ASR detects and processes speech as a signal, reducing background noise and converting them into phonemes. This chain of phonemes is analysed in a sequence and with the assistance of natural language understanding (NLU) software [11], ASR is able to use statistical probability to craft whole words and differentiate homophones.

2.1.2 Text-to-Intent

Text obtained from ASR can be taken out of context. Using natural language understanding, (NLU) voice assistants can understand the intent behind speech, differentiating between phrases such as 'travel 4 miles' and 'buy a present for miles.' This is accomplished through machine learning from historic interactions and associating certain words/phrases with certain contexts e.g. 'What's it like outside?' may refer to the weather outdoors. A useful infographic illustrating how NLU and ASR operate can be found at [12]. This combination of ASR and NLU allows for fluid interactions and intelligent conversion of speech-to-text.

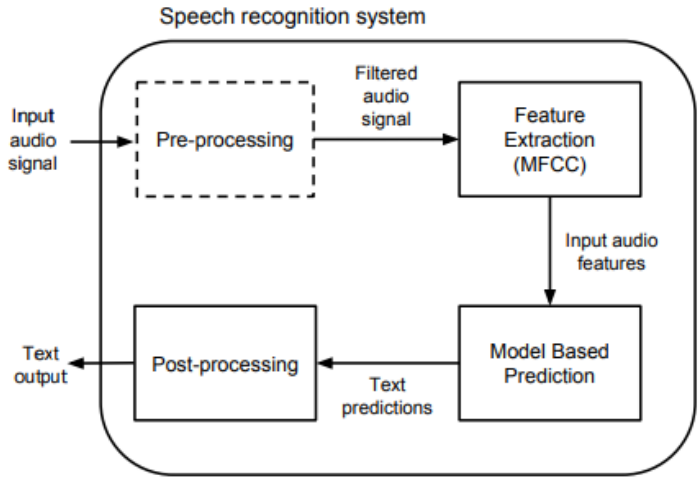

Figure 1: Overview of speech recognition systems [6].

2.1.3 Intent-to-Action

The final step takes the meaning of the text and converts this into real action. Voice assistants use text-to-speech synthesizers to acknowledge commands or reply to questions. Through machine learning, new words/phrases are learnt and remembered, priming voice assistants to understand specific mannerisms of the user [13].

2.2 Using Hidden Voice Commands to Attack

Voice command attacks achieve success by abusing the difference between human and machine speech recognition. Speech recognition systems extract features from audio using Mel-frequency cepstral coefficients (MFCC) but are unable to differentiate between human speech and obfuscated audio alone [6]. Hidden voice commands should satisfy two criteria: have sufficient features so that speech recognition systems can interpret the audio while also being unidentifiable by humans. Vaidya et. al. [6] created an audio mangler which extracts acoustic features from audio and uses an inverse MFCC to produce mangled audio. Producing the attack can be tedious, however once successful, attacks can be applied to a variety of systems which use MFCC. Figure 2 demonstrates the audio mangling process, including the speech recognition figure from figure 1.

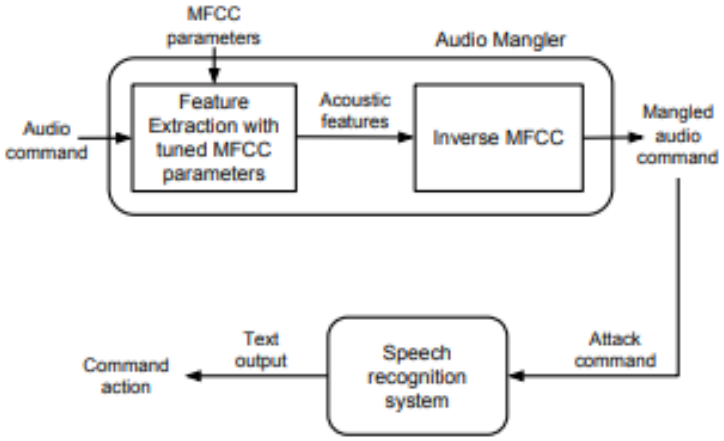

Figure 2: Audio Mangling Process [6].

Carlini et. al. [14] used this method to generate mangled audio commands for "Ok Google", "Turn on Airplane Mode" and "Call 911". Human comprehension was evaluated using Amazon Mechanical Turk, a service which pays human workers to complete tasks. Workers were asked to transcribe the aforementioned commands, given as audio files, with some workers receiving normal commands and others receiving mangled audio commands.

	Ok Google		**Turn on airplane mode**		**Call 911**	
	Machine	*Human*	*Machine*	*Human*	*Machine*	*Human*
Normal	90% (36/40)	89% (356/400)	75% (30/40)	69% (315/456)	90% (36/40)	87% (283/324)
Obfuscated	95% (38/40)	22% (86/376)	45% (18/40)	24% (109/444)	40% (16/40)	94% (246/260)

Table 1: Human and machine comprehension of normal and obfuscated audio commands [14].

Another example of this includes the Dolphin Attack, which injects voice commands completely inaudible to humans using ultrasonic carriers (frequency of 20 kHz). These attacks are created by the modulation of a low-frequency voice signal (baseband) on ultrasonic carriers, which are transmitted to speech recognition systems. Speech recognition systems demodulate these signals and output text. The Dolphin Attack was successful in attacking 16 devices which used voice assistants, including the iPhone 7 Plus, Apple Watch, Samsung Galaxy S6 Edge and Amazon Echo at a distance of a few feet with 55dB of background noise. [15].

2.3 Defence Mechanisms

There are currently no dedicated defensive measures against hidden voice commands. Recent focus has been on biometric authentication; however, emotion and identity characteristics are still a grey area [16]. Carlini et. al. [14] suggests three distinct types of defences: (1) Defences that notify, (2) Defences that challenge, and (3) Defences that detect and prevent.

Defences that notify include any defence method which notifies users of threats and prompt action to be taken. For example, beeping or buzzing whenever a device receives voice commands. Problems with this defence include being easy to mask or users ignoring notifications out of habit. Sunshine et. al. [17] has demonstrated that warnings which are easy to ignore such as beeps or small LED indicators are ineffective.

Defences that challenge include defence methods which seek user confirmation before execution. Such methods are effective, as they require attention to complete. An example of this is Google's reCAPTCHA [18]. Issues arise with reCAPTCHA due to audio CAPTCHA's being lenient due to usability issues. Due to this security flaw, inconvenience and cost inefficiency, active defences such as audio CAPTCHA's may not be feasible.

Defences that detect and prevent, such as speaker recognition and filters detect malicious attacks and prevent them from being executed. Such methods are effective against hidden voice commands while having minimal impact on the efficacy of normal voice commands but require additional research before implementation. Behaviour anomaly detection is within this area but not mentioned in [14].

2.4 User Profiling in Application

Google will send notifications or redirect to other pages when there is suspicious activity in an account. 'Suspicious activity' includes unfamiliar changes to critical security settings, unrecognized purchases, attempted sign-ins etc. [19]. Youtube uses user profiling in a different manner, providing video recommendations using a system which utilizes two neural networks [19]. An overview of this system is given in figure 3.

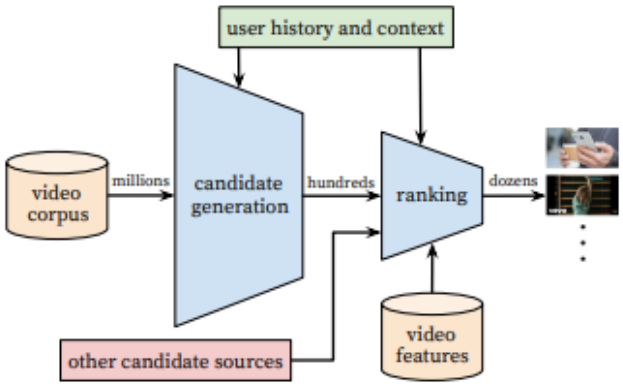

Figure 3: Overview of Youtube's video recommendation system [20].

From Youtube's large catalogue of videos, DNNs are used to generate recommendation candidates and rank them based on relevance. The DNNs factor in user history and context to funnel through candidate videos. This model of implementing DNNs which factor in user history to tailor recommendations can be applied to protecting devices from hidden voice commands, by using machine learning to understand how users interact with a device to prevent suspect behaviours. Haughay et. al. [21] states that devices can define users by interactions with devices.

3. Problem Identification

Voice assistants are becoming increasingly popular and the technology evolves, so are the hackers looking to exploit them for personal gain. To protect victims from these types of attacks, it is important to first understand how these attacks are generated and how they operate.

The main objectives of this project are to:

- Generate hidden voice commands which voice assistants will respond to and execute commands.
- Develop a framework for behavioural specification as a defence against hidden voice commands.
- Evaluate the effectiveness of the framework for securing devices.

4. Methodology

The project can be split into two main parts: generating hidden voice commands and developing a framework for behaviour specification. After generating the voice command, we can focus on voice command protection. The end goal is to establish a framework for behaviour specification, a key part of developing an application. To ensure that the trajectory of the project is in the right direction, consistent communication with the thesis supervisor, Jin Hong, is paramount.

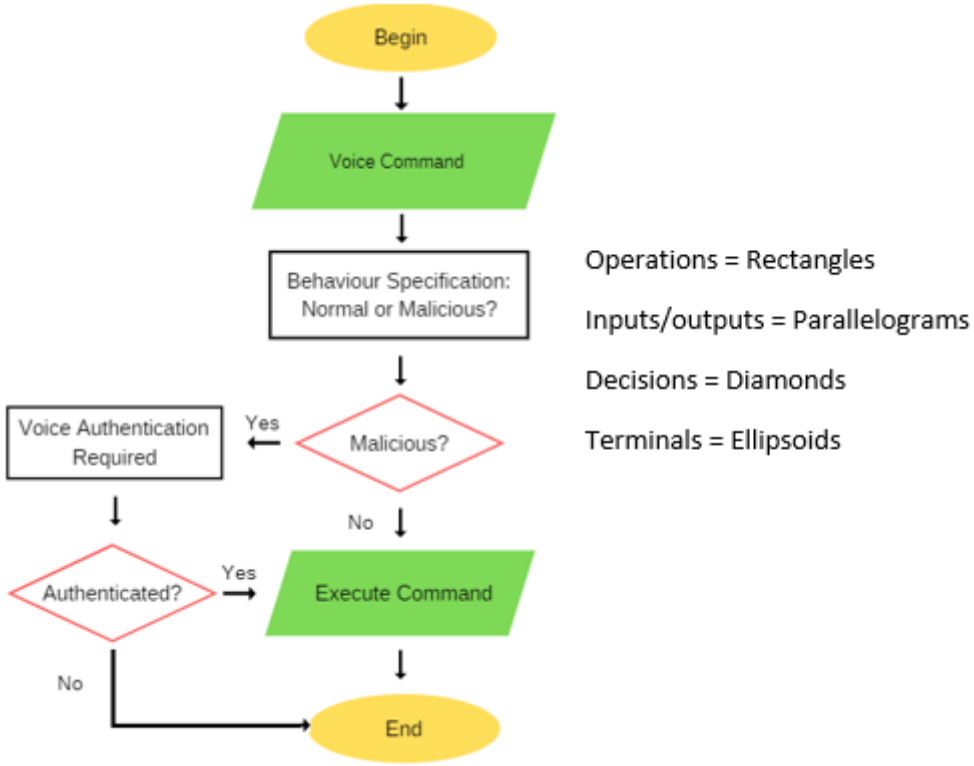

Figure 4: Malicious voice command detection flowchart

4.1 Hidden Voice Command Generation

4.1.1 Mangling Audio

The attack will be generated using MATLAB as per D. Ellis [22]. This method targets speech recognition systems which utilize MFCC for feature extraction and can also be modified to attack systems using Relative Spectral Transform – Perceptual Linear Prediction (RASTA-PLP) for feature extraction. However, we will focus on generating a signal to attack MFCC-using systems using the simplest approach possible. This is similar to the audio mangling method in [6], which extracts the acoustic features of audio and uses a code to invert MFCC. A list of the codes to be used to implement this can be found in the [22]. The code is capable of calculating the 20^{th} order MFCC features, giving us the desired mangled audio.

4.1.2 Attacking Speech Recognition Systems

The mangled audio may require some fine-tuning before voice assistants register them as commands. Using Google's Cloud Speech-to-text program, we are able to modify the audio until it is recognized as a command [23]. A Samsung S7 Edge smartphone will be used as a microphone, Google Assistant being the voice assistant to be attacked. To ensure reliability, some colleagues have offered use of a Samsung S10 edge and an iPhone X which use the voice assistants Bixby and Siri, respectively. The signal will be injected from a distance of a few metres via an Apple MacBook Air's speaker.

4.2 Malicious Voice Command Detection

4.2.1 Behaviour Specification

Behaviour specification is a passive defence integrated into devices which aims to categorize commands as normal or malicious based on its frequency of use. This is similar to how email services and banking applications alert users of suspicious behaviour. This is achieved through machines logging their user's habits, including websites they visit, location, financial habits etc. Commands which are frequently executed are flagged as normal, while less frequently used commands may require authentication. Potentially malicious commands will be suspended and require authentication to proceed.

Usage patterns will be generated by using voice assistants in daily life with a Samsung S7 Edge using Google Assistant. Google My Activity [24] will log voice command activity, providing conversation history and timestamps. Google Maps Timeline provides location history, allowing us to know approximately where we were when we issued voice commands. These services will provide us with the information necessary to determine the authenticity of the commands executed.

4.2.2 Decision Model

Threats will be determined using a probabilistic model, taking into account usage patterns, usage correlation and context (time and location) similar to an approach in [25] e.g. voice commands are typically used in the house (location of house mapped by general usage) and mainly during hours when users are awake. This model will be simulated on Tensorflow [26], where an input command is categorized by machine learning probability. An example of this in action would be sending an obviously malicious command with output: Probability of 'send $5000 to bank account in India' command at 4:00 am being malicious is YES, with probability 0.99 and NO, with probability 0.01.

4.2.3 Voice Authentication

Voice authentication is an active defence where users will be prompted to provide voice authentication when a system deems behaviour as malicious. The general idea behind authentication is ensuring that users are fully aware of commands which may threaten their security/safety. Implementing voice authentication after behaviour specification ensures that voice assistants can operate without constant interruption. This area of defence will be explored further by a colleague focussing on the same topic, Rui Lin.

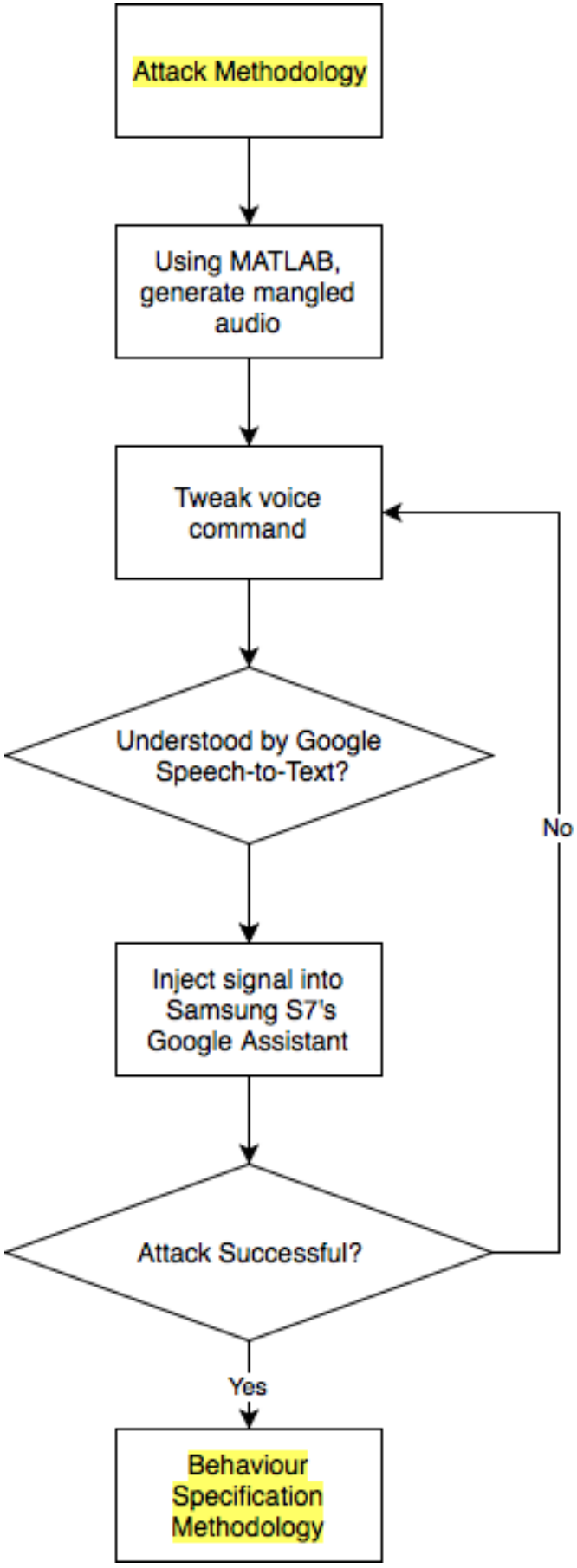

Figure 5: Generating the Attack

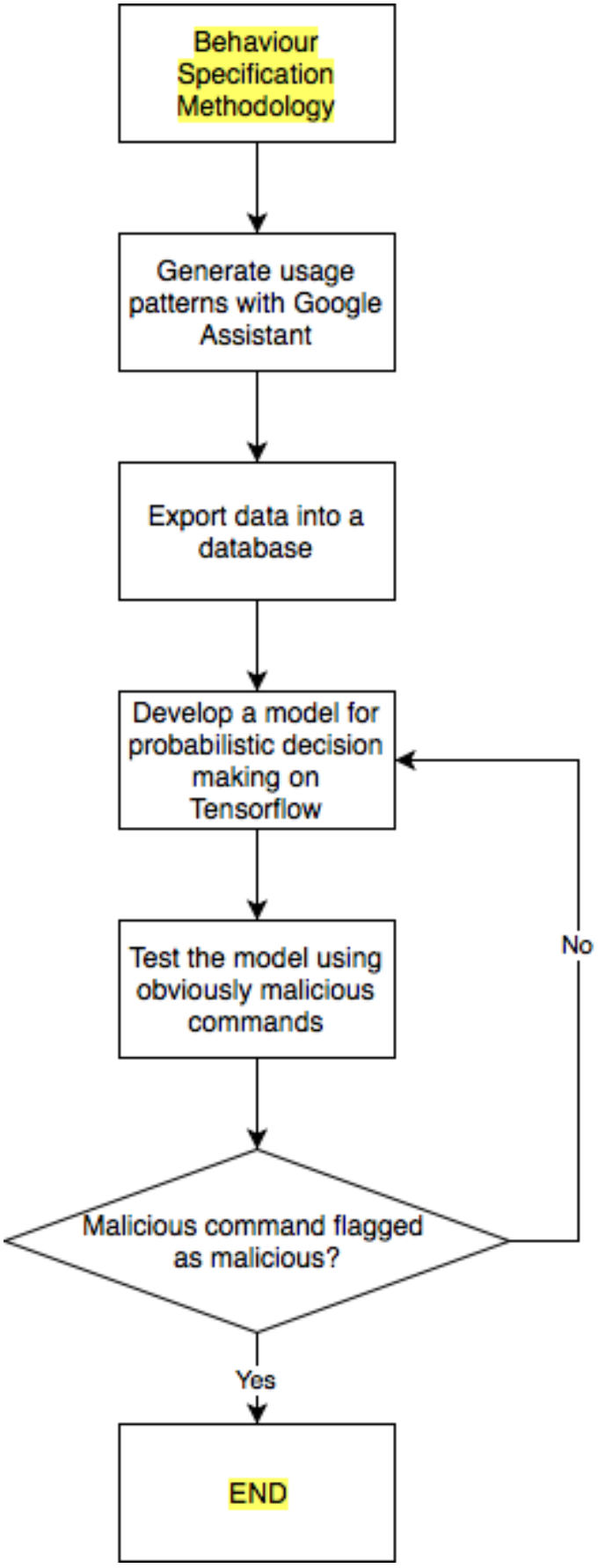

Figure 6: Behaviour specification framework

5. Timeline

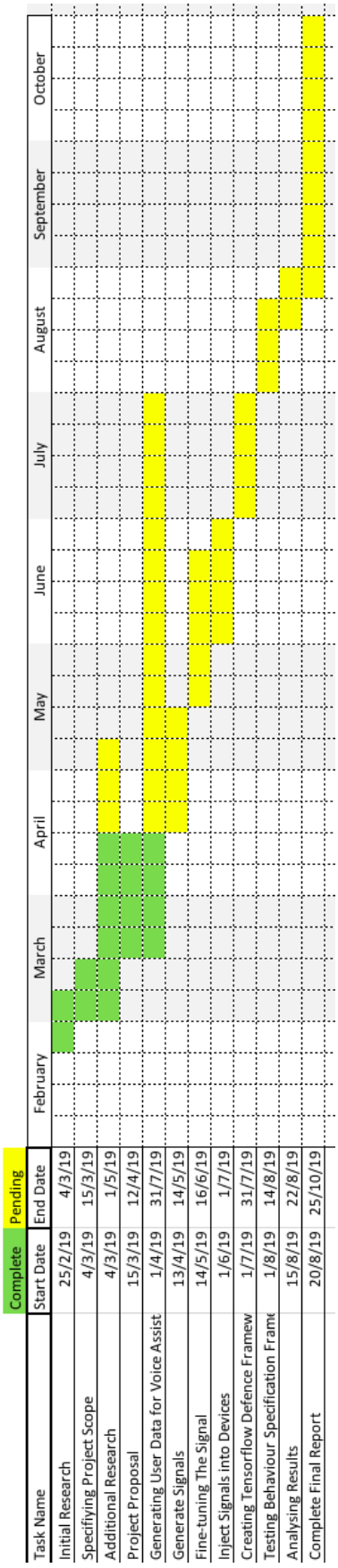

Figure 7: Thesis Gantt Chart

6. References

[1] M. B. Hoy, "Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants," *Medical Services Quarterly,* vol. 37, no. 1, January, 2018. [Online serial]. Available: https://www.tandfonline.com/doi/full/10.1080/02763869.2018.1404391 [Accessed March 15, 2019].

[2] B. Kinsella, "U.S. Smart Speaker Ownership Rises 40% in 2018 to 66.4 Million and Amazon Echo Maintains Market Share Lead Says New Report from Voicebot," *voicebot.ai*, Mar. 7, 2019. [Online]. Available: https://voicebot.ai/2019/03/07/u-s-smart-speaker-ownership-rises-40-in-2018-to-66-4-million-and-amazon-echo-maintains-market-share-lead-says-new-report-from-voicebot/. [Accessed March 16, 2019].

[3] S. Smith, "Digital Voice Assistants in Use to Triple to 8 Billion by 2023, Driven by Smart Home Devices," *Juniper Research*, Feb. 12, 2018. [Online]. Available: https://www.juniperresearch.com/press/press-releases/digital-voice-assistants-in-use-to-triple. [Accessed March 16, 2019].

[4] A. C. Clarke, "2001: A Space Odyssey," *United Kingdom: Hutchinsons & Co,* 1968.

[5] B. Bosker, "SIRI RISING: The Inside Story of Siri's Origins – And Why She Could Overshadow the iPhone," *The Huffington Post*, para. 35, Jan. 22, 2013. [Online]. Available: https://www.huffingtonpost.com.au/2013/01/22/siri-do-engine-apple-iphone_n_2499165.html. [Accessed March 16, 2019].

[6] T. Vaidya, Y. Zhang, M. Sherr, and C. Shields, "Cocaine Noodles: Exploiting the Gap between Human and Machine Speech Recognition," *9th USENIX Workshop on Offensive Technologies (WOOT)*, August, 2015. [Online]. Available: https://security.cs.georgetown.edu/~msherr/papers/cocaine-noodles.pdf. [Accessed March 08, 2019].

[7] M. E. Khan and F. Khan, "A Comparative Study of White Box, Black Box and Grey Box Testing Techniques," *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 6, June, 2012 Available: https://www.researchgate.net/publication/270554162_A_Comparative_Study_of_White_Box_Black_Box_and_Grey_Box_Testing_Techniques [Accessed March 20, 2019].

[8] Vivek Sharma, "How do digital assistants (e.g. Alexa, Siri) work?" USC Marshall, entry posted Oct. 17, 2017. [Online]. Available: https://www.marshall.usc.edu/blog/how-do-digital-voice-assistants-eg-alexa-siri-work. [Accessed March 24, 2019].

[9] D. A. Reynolds, "An overview of automatic speaker recognition technology," *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, pp. IV-4072-IV-4075. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5745552&tag=1. [Accessed March 24, 2019].

[10] Amazon inc., "ASR: A Foundation for Voice Technology," 2019. [Online]. Available: https://developer.amazon.com/alexa-skills-kit/asr. [Accessed March 27, 2019].

[11] Amazon inc., "What is Natural Language Understanding (NLU)?" 2019. [Online]. Available: https://developer.amazon.com/alexa-skills-kit/nlu. [Accessed March 27, 2019].

[12] M. Zajechowski, "Automatic Speech Recognition (ASR) Software – An introduction," *Usability Geek*, Dec. 29, 2014. [Online]. Available: https://usabilitygeek.com/automatic-speech-recognition-asr-software-an-introduction/ [Accessed March 27, 2019].

[13] R. Sepe, Jr., “Voice actuation with contextual learning for intelligent machine control,” U.S. Patent 5 212 821, May 18, 1993.

[14] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner and W. Zhou, “Hidden Voice Commands,” *25th USENIX Security Symposium (Security)*, August, 2016. [Online]. Available: https://security.cs.georgetown.edu/~tavish/hvc_usenix.pdf [Accessed March 07, 2019].

[15] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang and W. Xu, “DolphinAttack: Inaudible Voice Commands,” *2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 103-117. [Online]. Available: https://arxiv.org/pdf/1708.09537.pdf. [Accessed March 17, 2019].

[16] W. Yutai, L. Bo, J. Xiaoqing, L. Feng and W. Lihao, "Speaker recognition based on dynamic MFCC parameters," *2009 International Conference on Image Analysis and Signal Processing*, Taizhou, 2009, pp. 406-409. [Accessed 2 Apr. 2019].

[17] J. Sunshine, S. Egelman H. Almuhimedi, N. Atri and L. F. Cranor, “Crying Wolf: An Empirical Study of SSL Warning Effectiveness,” *18th USENIX Security Symposium,* 2009, pp. 399-416.

[18] Google.com, “reCAPTCHA,” 2019. [Online]. Available: https://www.google.com/recaptcha/intro/v3.html [Accessed 2 Apr. 2019].

[19] Google.com, “Detecting suspicious activity on your account,” 2019. [Online]. Available: https://support.google.com/accounts/answer/140921?hl=en [Accessed 2 Apr. 2019].

[20] P. Covington, J. Adams and E. Sargin, “Deep Neural Networks for Youtube Recommendations,” *10th ACM Conference on Recommender Systems (RecSys ’16).* Pp. 191-198. [Online]. Available: https://storage.googleapis.com/pub-tools-public-publication-data/pdf/45530.pdf. [Accessed 5 Apr. 2019].

[21] A. P. Haughay, M. Hill, “User Profiling for Selecting User Specific Voice Input Processing Information,” U.S. Patent 8 682 667 B2, Mar. 25, 2014. [Accessed 5 Apr. 2019].

[22] D. Ellis, “PLP and RASTA (and MFCC, and inversion) in Matlab using melfcc.m and invmelfcc.m,” *Columbia University*, [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/. [Accessed 4 Apr. 2019].

[23] Google.com, “Cloud Speech-to-Text,” 2019. [Online]. Available: https://cloud.google.com/speech-to-text/ [Accessed 4 Apr. 2019].

[24] Google.com, “Mr Activity,” 2019. [Online]. Available: http://myactivity.google.com/myactivity [Accessed 11 Apr. 2019].

[25] Lu, E.HC. & Yang, YW, “Mining mobile application usage pattern for demand prediction by considering spatial and temporal relations,” *Geoinformatica* (2018) 22: 693. [Online]. Available: https://doi.org/10.1007/s10707-018-0322-9 [Accessed 11 Apr. 2019].

[26] C. Davidson-Pilon, “Probabilistic Programming and Bayesian Methods for Hackers,” *Jupyter,* (2019). [Online]. Available: https://nbviewer.jupyter.org/github/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/tree/master/ [Accessed 11 Apr. 2019].