Clark High School

Preventing Caller ID Spoofing

Develop an application to prevent Caller ID spoofing, a type of identity theft

Kailash Subramanian

Computer Science / Mathematics, Engineering

19 January 2017

Acknowledgements

This study was impacted by the assistance of Riyaz Sikora, a professor and expert in the field of data mining, computing, and machine learning.

Contents

1. Introduction

Millions of people are affected by identity theft, and with the advent of technology, this is quickly becoming a major concern for users. Caller ID spoofing is a form of identity theft in which the caller ID, the number displayed on a phone through an incoming number, has been modified by a spoofer to appear as if the call is trustworthy. With the arrival of new types of technology like *Asterisk* or *FreeSWITCH*, it is easy for anyone to create a fake call to masquerade as a trustworthy caller, for example, from a bank, the government, and entice the user to give away personal information, resulting in identity theft.

This means that spoofers can masquerade as trustworthy banks, friends, or family members, and they can steal private information from the intended recipient. Alderman describes that spoofers have many methods of doing this, especially under the claim of a critical situation.[1] Spoofers can convince the user that he or she is under a critical and urgent situation in an attempt coax private information from them.

Unfortunately, little research has been conducted on this topic thus far. Although some mobile applications, such as Trapcall, help prevent scammed calls, they cannot prevent spoofed calls because the calls themselves appear to be trustworthy.[2] Another difficulty is that spoofed calls usually cannot be traced.[3] The initial solution involved blocking and redialing every single call that may pose a potential threat to identity theft; however, this ruthless attempt to block spoofed calls can be annoying to users. Although it may protect users from spoofed calls, it is not the most efficient or user-friendly solution.

This study involves a unique approach to this problem: voice analysis. Using data mining and voice analysis, it is possible to determine the legitimacy of a call and save lives of millions worldwide. The proposed solution involves an algorithm which determines the best course of action for an incoming call. This can involve redialing the number, voice frequency analysis, word usage analysis, or completely blocking the number, to prevent Caller ID spoofing in a strategic manner.

This project can help save the lives of millions of people around the world who fall to scammers that steal private information.

2. Proposed Solution/Method

This engineering project included the development of a mobile Android application.[4] The proposed solution involves the usage of an algorithm to determine the best course of action for an incoming call. A complete flowchart of this algorithm can be found on Section 2.1.

Voice frequency analysis is a possible course of action in which the program determines the frequency of the caller's voice. The frequency is compared with existing data in a database collected from prior calls. If the voice frequency conflicts with existing data, the call will be

blocked. However, if the caller has a valid voice frequency, then the program will proceed to word type analysis. This involves speech recognition. The words will be extracted from speech and processed in another algorithm, which determines the possible location of the caller based on the dialect of English and types of words they used. This is compared with the area code given in the phone number. If this test fails, then the call is blocked.

If the call is claiming to be from a trustworthy organization, such as a governmental number, the call will be placed on hold. The user will be asked to dial back from the keypad. If the second call passes through, then the initial call had been spoofed, since the initial call is claiming to be in a call. If the second call does not pass through, then the call may be valid. Word type analysis will be performed after this test. After five minutes, the call will automatically be cut, and subsequent calls from this number will be blocked.

If the call is foreign (not from a trustworthy organization), the program will perform word type analysis. After five minutes, the call will automatically be cut, and subsequent calls from this number will be blocked.

In an outgoing call, the call will not be spoofed because the user has created the call. The program will record the number that was called and determine the voice frequency of the intended recipient of the outgoing call since this data will be used in voice frequency analysis. The information will be stored in a database.

For additional security, the program will also allow the user to select numbers to block. Calls from these numbers will automatically be cut, and the user will be asked to call back if they wish.

2.1 Flowchart

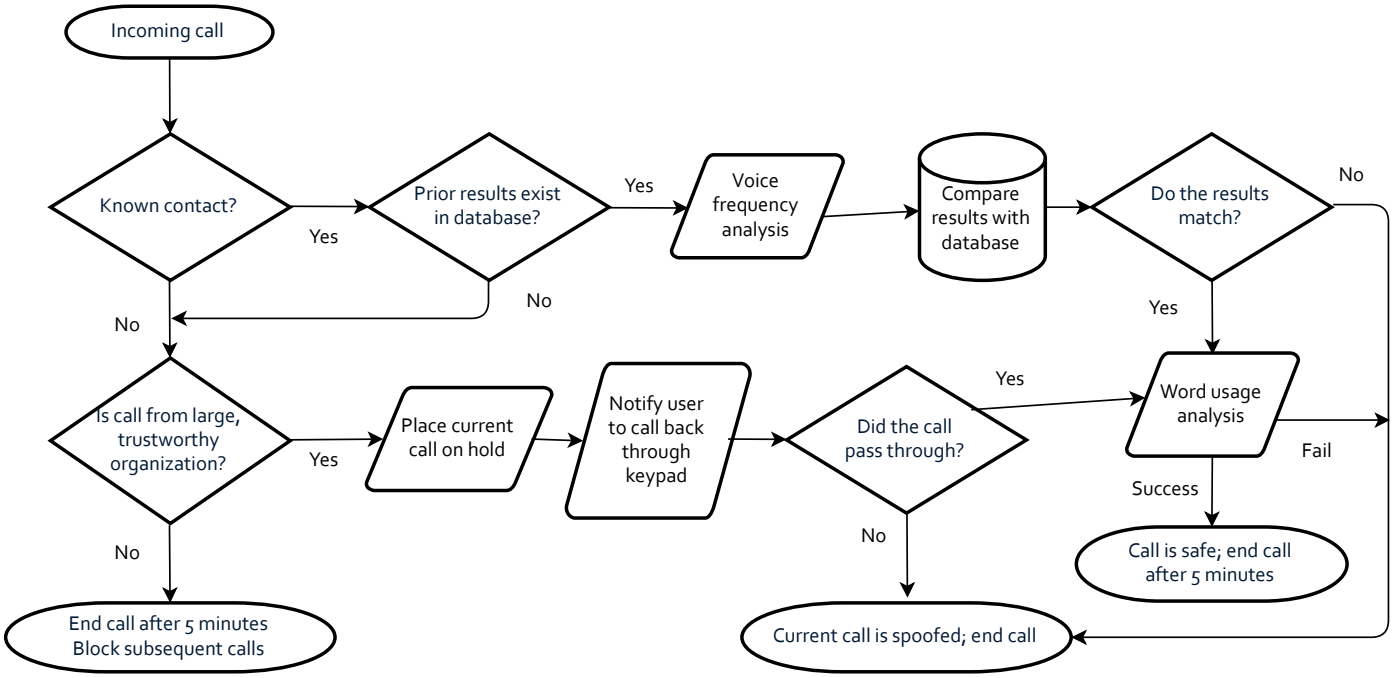

Figure 1: A compendious flowchart depicting the program's course of action for an incoming call

2.2 Voice Frequency Analysis

Voice frequency is a feature of the human voice that can be used to distinguish humans from each other.[5] This can be useful in determining whether the caller is indeed who they claim to be. Voice frequency analysis is used if the caller's frequency information has previously been recorded and stored in a database.

In an outgoing call, the program will determine the voice frequency of the intended receiver. Since the call will not be spoofed, it is safe to record the voice frequency in a database as it will be legitimate. In an incoming call, the program will look up the number and its corresponding voice frequency in the database. If the voice frequency is within a 10 Hz range, then the voice is most likely legitimate, and it matches the phone number.

2.2.1 Voice Frequency Algorithms

In Java, voice frequency can be found in a multitude of ways. After testing a variety of algorithms, including the Discrete Fourier Transform (DFT) and the Autocorrelation Function, the Average Magnitude Difference Function (AMDF) seemed to be most successful as it used less computationally expensive mathematical functions.

2.3 Word Usage Analysis

English has changed rapidly over the past few hundred years, and it has shaped itself to become part of the identity of those in different regions. In the U.S., for example, there is a large

distinction between the dialects of English, ranging from the Gulf Southern dialect to the Rocky Mountain dialect.

Thus, it can be possible to determine the location of a caller based on the types of words they use. In this algorithm, which, for now, only considers the United States, 8 regions have been selected for distinctions in English word usage; these dialects consist of Pacific Northwest, Pacific Southwest, Rocky Mountain, Gulf Southern, Coastal Southern, Midland, Upper Midwestern, and Northeast.

Word usage analysis is used if voice frequency analysis was successful or if a call from a trustworthy organization seems to be legitimate. Word usage analysis ensures the security and legitimacy of a call through a second test.

The algorithm simply uses a variety of key words that show distinctions between these regions using data from Harvard's Dialect Survey as well as Grieve's World Mapper. Although a variety of words were used, the algorithm also focused on a set of words relating to money.

This technique is advantageous in that it can predict the location of the caller without relying on tracing the call, which one source says is nearly impossible if the call is spoofed.[3]

2.4 Calls from Large, Trustworthy Organizations

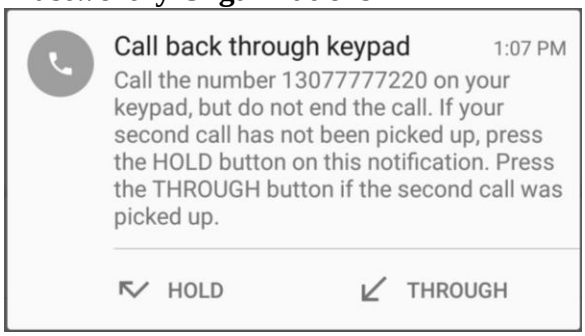

Figure 2: Android notification asking the user to call back to a number

Spoofers are likely to impersonate a person from a large organization, such as the government, to appear trustworthy and obtain money from their victims. In this case, the user should answer the call. The program will send a notification to the user, asking them to call back the number.

Once the user has called back via the keypad, they can simply tap the button representing the state of the call: passed through or not been picked up.

If a second call had been passed through, that means that that the initial call had been spoofed: it is not possible for two calls to be ongoing at once – the initial call must have been from a different caller, thus being fake.

On the other hand, if the second call was not picked up, it shows that the initial call was still ongoing. Thus, the initial call was most likely legitimate.

2.5 Android Application

The user interface for developing the Android Application followed Google's material design guidelines. The primary color was teal, while the secondary color was green. The Android application was tested on the Nexus 5 emulator, running API 23 (Android 6.0 Marshmallow).

Since the Android emulator does not support audio input via the computer's microphone, the code for voice analysis had been created on a separate IDE and run on the computer as a standalone application that utilizes the computer's microphone.

3. Materials

Minimum hardware requirements

- A computer running Windows 7 or later, Mac OS X 10.8.5 or higher (up to 10.11.14), or a Linux GNOME or KDE desktop, with a functional microphone
- 4 GB of RAM

Minimum software requirements

- Android Studio IDE
- Eclipse IDE
- Vector graphics editor, such as Inkscape
- JDK 7 or later

Human participants for experimentation

- Four human participants (including student researcher)
    - Two male participants: one between 10 and 15 years of age, one adult
    - Two female participants: one between 10 and 15 years of age, one adult

4. Procedure

1) Download and set up Android Studio. Develop the application and test it through the emulator provided by the IDE.
2) Follow the Google material design specifications to make sure the application stays consistent with the conventions normally used while developing an Android application.
3) Since the Android emulator cannot input sound, develop the code relating to voice frequency analysis and word usage analysis on a separate Java project using Eclipse, a different IDE.
4) Input 12 recorded calls of equal length into the microphone. There are 3 sentences spoken for each of the four different people. This includes the experimenter and three other participants. In total, there should be two male participants and two female participants.

Let the application determine the frequencies of the voice in the recordings and store it in the database.

5) Input 9 recorded calls of equal length into the microphone, where 3 sentences are spoken from each phone number. 4 of these calls are spoofed with a different voice type. The remaining 5 of these calls are calls that are legitimate.
6) For step 5, record the time taken for the application to determine the voice frequency and the legitimacy of the call based on the phone number. Record the time taken for each prediction as well as the confidence level of the program.
7) Repeat steps 5 and 6.
8) Make changes to the code until the desired accuracy of at least 99% and latency time of an average of under 1 second are reached.
9) Redesign and retest until optimal conditions are met.

5. Experimentation and Results

5.1 Initial Plan for Experimentation

The initial means of experimentation differs from the final means of experimentation, which can be found on Section 5.2.

To test the application, 20 phone numbers on pre-recorded calls was stored in the database, along with the corresponding voice data of each call. Create 40 recorded calls, where there are 2 calls from each phone number, and 20 of these calls are spoofed with a different voice type.

The time taken for the application, to determine if it was spoofed or not, is recorded, along with the accuracy of the prediction. Redesign as needed, and repeat the process until the desired accuracy and latency times are achieved.

5.2 Final Form of Experimentation

The means of experimentation had to be changed due to the usage of an emulator. Although it is possible to emulate an incoming call, it is impossible to send voice data along with an incoming call. In addition, the emulator cannot receive audio input from the computer's microphone.

For these reasons, voice analysis had to be developed on a different IDE (the Eclipse IDE was used in this study) and tested on a standalone JavaFX application. The means of experimentation also had to be altered and is described below.

Human participants will speak into a microphone to simulate a normal call. Each participant will speak to create 6 simulated calls, all of which are the same length. The 6 calls will be recorded into a microphone, then it will be played into the computer's microphone to ensure that all calls are of equal length.

3 simulated calls will be used for *learning*. 3 different simulated calls will be used for *testing* (experimentation). *Learning* refers to the program learning the voice type of the call.

Sentences used for learning: (exactly 8 seconds each)

1. The North Wind and the Sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak.
2. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other.
3. Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveler fold his cloak around him.

Sentences used for testing: (exactly 5 seconds each)

1. And at last, the North Wind gave up the attempt.
2. Then the Sun shined out warmly, and immediately the traveler took off his cloak.
3. And so the North Wind was obliged to confess that the Sun was the stronger of the two.

In experimentation, only the testing calls are taken into consideration for measuring latency times and accuracy. The control group was a middle C (261.6 Hz) played on a piano, which was identified as "Person Z."

5.3 Results (Data)

The data obtained from experimentation is shown below.

Prediction of Call Legitimacy Based on Voice Frequencies

Type of Call	Call Number & Trial Number	Prediction & Actual Caller	Accuracy of Prediction
Middle C on a Piano (*Control Group*)	0	99.4% confident: Z Actual: Z	Correct
Spoofed Call	1, Trial 1	97.2% confident: D Actual: B	Incorrect
	1, Trial 2	98.9% confident: B Actual: B	Correct
	2, Trial 1	94.4% confident: B Actual: B	Correct
	2, Trial 2	97.8% confident: D	Incorrect

		Actual: B	
	3, Trial 1	99.7% confident: C Actual: C	Correct
	3, Trial 2	92.4% confident: C Actual: C	Correct
	4, Trial 1	98.0% confident: C Actual: C	Correct
	4, Trial 2	99.8% confident: C Actual: C	Correct
Legitimate Call	5, Trial 1	93.3% confident: D Actual: D	Correct
	5, Trial 2	94.4% confident: D Actual: D	Correct
	6, Trial 1	94.4% confident: D Actual: D	Correct
	6, Trial 2	95.6% confident: D Actual: D	Correct
	7, Trial 1	95.1% confident: D Actual: D	Correct
	7, Trial 2	96.6% confident: B Actual: D	Incorrect
	8, Trial 1	97.8% confident: B Actual: D	Incorrect
	8, Trial 2	99.5% confident: D Actual: D	Correct
	9, Trial 1	99.7% confident: D Actual: D	Correct
	9, Trial 2	91.1% confident: D Actual: D	Correct
Average Results		**Average Confidence ≈ 96.43%**	**14/18 correct 77.778% accurate**

Figure 3: Prediction of Call Legitimacy Based on Voice Frequency

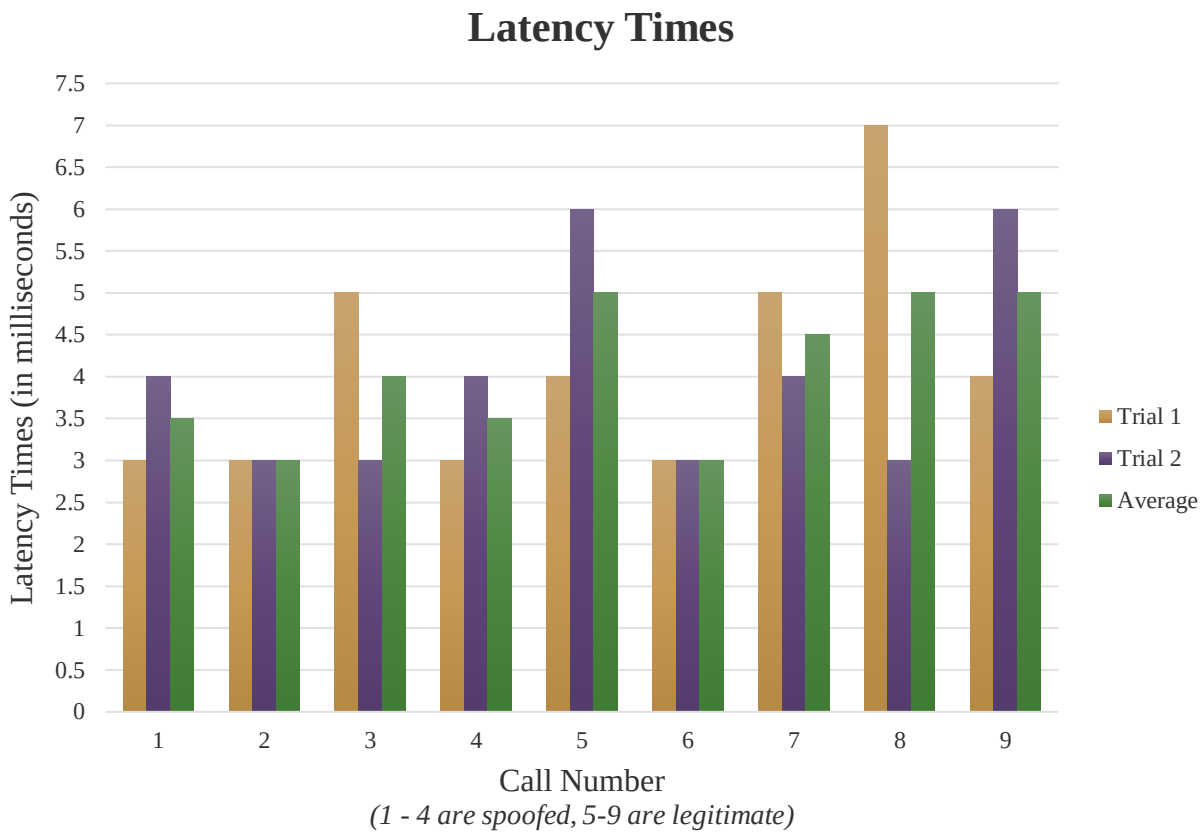

Figure 4: Latency times. Average latency time was 4.055555556 milliseconds.

5.4 Results (Data Analysis)

The data shows that the prototype was very successful. It could differentiate between common voice types and ages, since subjects A and B were female, and C and D were male. The prototype was successful in about 78% of the trials. This can be improved by implementing a more complicated algorithm involving complex voice frequency analysis, in regards to the types of words used by the caller and the timbre of the voice.

The prototype was also very confident. In fact, confidence levels were over 90% for all trials, and the average confidence was about 96%. This was expected because the prototype was only tested with four voices in the database. Once the number of items in the database increases, linear searching will soon become slow. In the future, binary searching will be implemented for a faster time. Confidence levels should also what naturally decrease.

6. Conclusion

Caller ID spoofing, the masquerading of a trustworthy caller via the alteration of Caller ID, is used by fakers to obtain private information from their victims. A spoofer might act like a person's account is threatened, so the victim can be willing to give away private information to protect their account. Caller ID spoofing affects millions worldwide.

The purpose of this project is to develop an Android mobile application that can detect spoofed calls. The design criteria was to determine if a call was spoofed via data mining, detect a spoofed call and have a latency time of almost 0 milliseconds, predict the type of caller based on prior call recordings, and notify the user if a call is spoofed.

The proposed solution was to use voice frequency analysis and data mining if the user has spoken to the caller in the past, place the current call on hold and have the user call back the number from the keypad in the caller is from a large, trustworthy business organization, and end the call in 5 minutes if the call is foreign, and possibly not from a trustworthy caller.

To test the prototype, four participants, subjects A, B, C, and D, each spoke 3 sentences for the program to learn the voice frequency. Next, each subject spoke 3 sentences for experimentation, and the prototype was successful in determining a spoofed call, where the speaker didn't match the Caller ID, in 14 out of 18 calls. There were 9 total calls input into the program, randomly decided as legitimate or spoofed. Each call had two trials, in a total of 18 calls. The program was about 78% accurate and about 96% confident in its predictions. The average latency time was about 4.1 milliseconds.

The data generally supported the design criteria, but it can be improved to fit the proposed solution. The latency time can be reduced by the usage of binary searching through a database and the exploration of different algorithms, such as the FFT and the Goertzel algorithm. The accuracy rate can be increased through the usage of complex voice analysis, through the analysis of timbre and accents. A spoofer can utilize programs that make it possible to change one's voice, so this can work around my current prototype. The frequency of the human voice can also change if the person is sick, or has a cold. Average frequency detection may not be accurate if there are pauses in the call, and background noises also need to be considered.

The experiment can prevent Caller ID spoofing, a type of identity theft, and save the lives of people around the world from calls that attempt to obtain private information. Further studies include the detection of mood in voice recognition software and aiding humans psychologically based on the tone of one's voice. This can also be used by personal assistants to cheer up one's mood if they are melancholy.

7. Error Analysis

There are many programs that allow spoofers to change their voices. Thus, the prototype cannot detect changed voice frequencies to match the frequencies claimed in the Caller ID.

When a human is sick, for example, with the common cold, their voice frequency naturally becomes deeper.

If the call is short with many pauses, the detection of average frequency can decrease.

Background noises need to be considered as it can affect the average frequency. For example, if someone is singing in the background of a call, then the average frequency might be higher than normal.

8. Future Improvements for Prototype

Timbre is the mixture of different waves to create a sound. Two instruments that play the same note, for example, will sound differently even if they play at the same frequency and volume because they have different timbre. Analysis of timbre in the human voice can greatly increase distinction between voices for the program.

In the future, it is necessary to test word usage analysis as well. This can be done by generating sentences and inputting them into the program. The program will predict the caller's location within the United States. The results of this experimentation can be used for further refinery of the prototype.

Currently my program can predict the locations of people inside the United States. However, people in different parts of the world also use different types of words. A spoofed call can be detected if the caller is, for example, an overseas caller, since they may have a significantly different word choice than the expected accents from past calls or the area code, if the call is foreign. Thus, the prototype can be improved by extending the range of predictions to the entire world. Similarly, the prototype can be improved by detecting accents and tone of the voice as well.

Usage of other algorithms, such as the FFT, DTSTFT, and the Goertzel algorithm can be tested to attempt to minimize latency in detection of voice frequencies.

Usage of other algorithms, such as binary searching, can be used for searching through a database with a smaller latency time, instead of linear searching.

9. Future Applications & Future Research

9.1 Future Applications

Voice recognition can be used to prevent Caller ID spoofing, a type of identity theft that affects people worldwide.

9.2 Future Research

Research has shown that voice recognition in differentiating between different people is possible via voice frequency analysis. This can be extended to involve experiments in determining the mood of the caller as well as the location of a caller based on the accent and the types of words used. This is useful not only for preventing Caller ID spoofing, but even to block scammed calls, emails, and protecting identity theft.

Notes

[1] "Under the pretext of warning about an urgent situation," Alderman states that spoofers are likely to "coax you into revealing personal or account information, supposedly to verify their records" (8).

[2] In their article "Caller ID Spoofing: All You Need to Know," Trapcall states that it "is not possible to prevent oneself from receiving spoof calls" (25). Trapcall can only unmask "blocked, [calls with] no caller ID, unknown, restricted, or private calls" (sec. 13).

[3] Unless "in extreme circumstances involving law enforcement, caller ID spoofing can NOT be traced" ("Caller ID Spoofing: All You Need to Know", sec. 14).

[4] This project included developing both an Android application, which includes user interface, notifications, and the logic, as well as a JavaFX application, which includes algorithms for voice analysis. The code was uploaded on to GitHub (see *github.com/kaisubr/ monitor_android_studio* and *github.com/kaisubr/monitor_voice_analysis*).

[5] Human voices have a wide range of frequencies, and the normal "voice range is about 500 Hz to 2 kHz [where] low frequencies are vowels and bass [and] high frequencies are consonants" (Marshall 3).

Works Cited

Alderman, Jason. "Don't Get 'Spoofed' by Rogue Callers." *The Huffington Post*. The Huffington Post, 7 Oct. 2013. Web. 26 Sept. 2016.

"Caller ID Spoofing: All You Need To Know." Trapcall. Trapcall, 6 July 2016. Web. 25 Sept. 2016.

Dunham, Ken. "Phishing, SMishing, and Vishing." Mobile Malware Attacks and Defense. Burlington, MA: Elsevier, 2009. 135. Print.

"Introducing Hiya: A Better Phone Experience." Hiya, 26 Apr. 2016. Web. 24 Sept. 2016.

Marshall, Dave. "Human hearing and voice." Human hearing and voice. N.p., 4 Oct. 2001. Web. 25 Sept. 2016.

Poulsen, Kevin. "VoIP Hackers Gut Caller ID." The Register. N.p., 7 July 2004. Web. 26 Sept. 2016.

"Spoofing and Caller ID." Consumer Help Center. Federal Communications Commission, n.d. Web. 18 Sept. 2016.

Swoboda, Andrew. "How to Protect Yourself From Caller ID Spoofing." The State of Security. Tripwire, 20 Apr. 2015. Web. 22 Sept. 2016.

The New Truecaller – A Smarter Way to Make Calls. Perf. Rishit Jhunjhunwala and Svetlana Lekoska. Truecaller, 26 Apr. 2016. Web. 25 Sept. 2016.

"Voice Over Internet Protocol (VoIP)." Federal Communications Commission. Federal Communications Commission, n.d. Web. 18 Sept. 2016.

"What's Spoofing?" CenturyLink. N.p., n.d. Web. 25 Sept. 2016.

Statement on Outside Assistance – This form should be included as the last page of your research paper.

Texas Junior Science and Humanities Symposium

Name: Kailash Subramanian

Title of paper: Preventing Caller ID Spoofing

1) What steps led you to formulate your hypothesis? (Where did you get the idea for your research?) Please be specific.

Caller ID spoofing is a type of identity theft in which a caller masquerades a trustworthy identity and steals information from their victim. Last summer, my parents were victims of Caller ID spoofing. I suddenly became aware of the different types of identity thefts that affect unsuspecting people, and thus, the idea for my project emerged.

2) Where did you conduct the major part of your work? (i.e. home, school or other institutional setting, university lab, medical center, etc.)

The major part of the work was conducted at home.

3) If you worked in an institutional setting, did you work on your project as part of a team or group? If so, how large was the team, and who was on the team (students, adult researchers, etc.)? Describe your role on the team.

The work was conducted in an institutional setting.

4) Describe what parts of the research you did on your own and what parts where you received help. (i.e. literature search, hypothesis, experimental design, use of special equipment, gathering data, evaluation of data, statistical analysis, conclusions, and preparation of written report (abstract and/or paper).

I received help in designing a solution to prevent Caller ID spoofing. All other parts of the research, including literature search, gathering data, evaluation of data, conclusions, and preparation of the written report, was done on my own.

5) If this research is a continuation of an investigation that was previously submitted to a regional JSHS, describe how you have expanded your investigation.

This research is not a continuation of an investigation that was previously submitted to a regional JSHS.

Kailash Subramanian