Extracting Dialed Telephone Numbers from Unstructured Audio

Steven Presser
Presser Surveillance Solutions and Technologies
P.O. Box 1271
Lynnfield, MA, United States
spresser@presser.tech

Michael Walsh
Presser Surveillance Solutions and Technologies
P.O. Box 1271
Lynnfield, MA, United States
mwalsh@presser.tech

***Abstract*—presented is a novel method (the Call Contents Automatic Differentiator (CCAD)) for extracting dialed telephone numbers from unstructured audio without capturing audio content – including other dialed information. This technology fills a critical law enforcement need to determine the ultimate destination of a call, even when the call is routed through multiple redirectors. The basic methodology involves examining the timing between digits, as well as the volume of audio segments between digits. Despite its simplicity, this method was able to isolate and extract dialed telephone numbers with accuracy greater than 99% in the expected scenario and greater than 98% in worst-case scenarios. If expanded to work in real-world scenarios, CCAD could serve as a new and clearly-legal source of information for local, state and national law enforcement, as well as operating in national security cases.**

***Index Terms*—Surveillance, Telecommunications, Telephony, Pattern matching**

I. Introduction

In this paper, we describe CCAD, the Call Contents Automatic Differentiator, a naive system for determining if a sequence of digits dialed after a telephone call has connected is a telephone number (routing data) or other information (content). These digits are known as "Post-Cut-Through Dialing Digits" or PCTDD. The system runs with an expected accuracy of 99.4 percent and 98.3 percent in the worst case. Such a system enables differentiating so-called "envelope/routing information" (which may legally be collected without a warrant) from "content information" (which may not). This paper describes the algorithm used.

This technology fulfills a critical law enforcement need for access to digits dialed after a call connects. This need is demonstrated and discussed in [1]. In that case, the court determined that a national security need was significant enough that access to post-cut-though dialing digits could be allowed without minimization technologies. But the court made it clear that this was unusual and that they would be more comfortable if minimization technologies were available. Additionally, the court made clear that the decision is limited to national security cases.

The goal of CCAD is to differentiate between dialed envelope information and other dialed information. Further, it attempts to do so in a robust, reliable, and real-time manner. CCAD should thus provide law enforcement access to these dialed digits in both national security and criminal contexts.

This problem has not previously been directly addressed in the literature. This paper is intended as an initial, laboratory quality solution. The expansion of this algorithm to real-world settings is left for future work.

We will first discuss the history and technology of the telephone, as well as relevant legal background. Then we will detail the algorithm itself. Next, we discuss the results. Finally, we discuss improvements which would enable the use of this algorithm outside a laboratory setting.

II. Background

In modern society, we often dial telephones but rarely think about what is required to connect a telephone call. This section explores this topic, as well as some telephonic and digital signals processing (DSP) history. Additionally, relevant legal background is included.

A. Telephone History

The network used to connect one telecommunications user to another is the PSTN (Public Switched Telephone Network). The first deployment of what would eventually become the PSTN was Bell Telephone Company's, in 1878 [2]. Dialing methods have evolved and adapted as the network has grown and as technology has advanced. Initially, operators were required to connect every call - manually plugging short lengths of cable to connect different "circuits" (essentially creating a point to point telephone line). Later, rotary dialing was introduced as a way to automate dialing and reduce the number of operators required. In 1960, the first paper on DTMF (Dual-Tone Multi-Frequency) dialing was published [3]. The first introduction of DTMF to the PSTN happened on November 18, 1963 [4]. With minor variation, DTMF has remained the standard since. Today, the DTMF standards are detailed in the International Telephone Union's recommendations Q.22 [5], Q.23 [6] and Q.24 [7]. Recently, much of the phone system has been digitized, but the user-facing interface (DTMF) has remained the baseline technology to interact with the telephone system.

B. DTMF

DTMF at its core is a set of eight tones (four high and four low) [6]. Each pair of tones (one from the high set, one from the low set) conveys one of the signals 0 through 9, A through D, the star and the octothorpe ("pound sign"). Though the signals A through D never made it into consumer use, they remain in the standard.

DTMF decoding software has been around since the origin of Digital Signals Processing (DSP). The oldest freely-available paper on implementing DTMF detection in software we can locate is from 1989 [8]. However, we are confident this is not the first software implementation – if nothing else, there would have been proprietary implementations. Paper [9] after paper [10] describing various implementations has followed, as have open-source implementations [11] [12] [13].

C. Voice Activation Detection

Our algorithm also performs Voice Activation Detection (VAD) – determining which parts of audio contain a person speaking and which contain noise. VAD is an area of ongoing research. However, this paper relies only on one of the many measures used in VAD [14] – Short Term Energy detection. Energy detection simply calculates the average energy of sections of audio and is the most obvious and most simple possible measure for whether or not there is speech in an audio stream. However, it performs extremely poorly if the environment is noisy or if sections of the speech are more quiet than others.

D. Legal Background

The legal grounding for capturing dialed digits but not the full content of a call goes back to before the telephone. In *Ex Parte Jackson* [15], the Supreme Court decided law enforcement was able to examine the outside of mail without a warrant, but that a warrant would be required to open it. This created a distinction between "envelope information" (*ie*: metadata or routing information) and "content" that persists today.

This became critical in *Smith v. Maryland* [16], where law enforcement installed a pen register (a device to record the destination of outgoing calls) to catch Smith making harassing telephone calls. The Supreme Court held that this was not a search, as Smith had no "reasonable expectation of privacy" in the phone numbers he dialed – because they had been voluntarily disclosed to a third party. This case, combined with the 1968 Omnibus Crime Control and Safe Streets Act [17] creates the legal framework for modern wiretaps and dialing data collection.

However, the status of digits dialed after a call initially connects is unclear. Some of these digits are clearly content and protected by a warrant (*eg*: credit card numbers, personal identification numbers). Other digits may be routing information (*eg*: final phone number to call when using a calling card service). Nor does the 1968 Omnibus Crime Control Act [17] provide guidance on the capture of PCTDD, as it predates the common use of systems which would require dialing after a call connects. The clearest answer comes from the Federal Communications Commission, which found in 1999 that "some digits dialed by a subject after connecting to a carrier other than the originating carrier are call-identifying information." [18] However, the FCC did not address how to extract only the digits which qualify as call-identifying information. In general, the domestic courts have declared that absent a technology to ensure that only call-identifying information is captured, no digits whatsoever may be captured.

The future of the "third party doctrine" underlying this framework is currently in doubt. In two recent location-tracking cases [19] [20], the Supreme Court has determined that merely revealing information to a third party is not enough to invoke the third party doctrine. It is unclear if and how this new line of reasoning applies to the legality of pen registers.

E. Law in the Real World

The impact of the legal limitations on the collection of envelope information are significant. Criminals, especially organized crime, are heavily dependent on telecommunications to coordinate their activity across multiple continents [21].

The value of these communications is not lost on law enforcement. "United States law enforcement agencies generally agree that electronic surveillance may be the most important and sophisticated investigative device available in the prevention, investigation, and prosecution of organized crime. In the world of drug trafficking, electronic surveillance is often the only method available to intercept communications between the drug kingpin and his highest officers within the crime enterprise." [22]

However, as always, criminals seek clever ways to evade detection. An entire suite of anti-surveillance methods exists int he criminal underworld, from the use of codewords to easily-discarded "burner" phones, including measures designed to evade pen registers. "Pre-paid calling cards also remain an effective alternative for criminals seeking to evade electronic surveillance since law enforcement can no longer intercept the "post cut-through" dialed digits [which are the ultimate destination of the call]" [23].

Thus, a technology which could separate PCTDD representing envelope information from PCTDD representing content would be of significant use to law enforcement, allowing the collection of a new type of information in a legally-valid manner. In turn, this would ensure the evidence gathered is admissible and could serve as the basis for convictions.

III. The Algorithm

CCAD uses two-stage process to determine which portions of the audio stream constitute envelope information. It assumes most relevant audio is going to be interaction with automated systems. Most automated systems use a prompt-and-response format. In this format, the system first prompts a user for information via a vocal prompt, then the user responds with the information via pressing buttons (which sends DTMF over the telephone line). The use of a prompt-and-response format

implies that there are sections of voice separating the DTMF-encoded responses and that digits can therefore be grouped where they are not separated by sections of voice or significant silence.

Stage 1 is the extraction of a "signal stream" from the audio, containing all DTMF signals and all separators. Stage 2 examines the signal stream using simple pattern-matching filters to determine what actually is envelope information. The final output of this methodology is any envelope information that was embedded in the audio stream. An overview of this process is provided in Fig. 1.

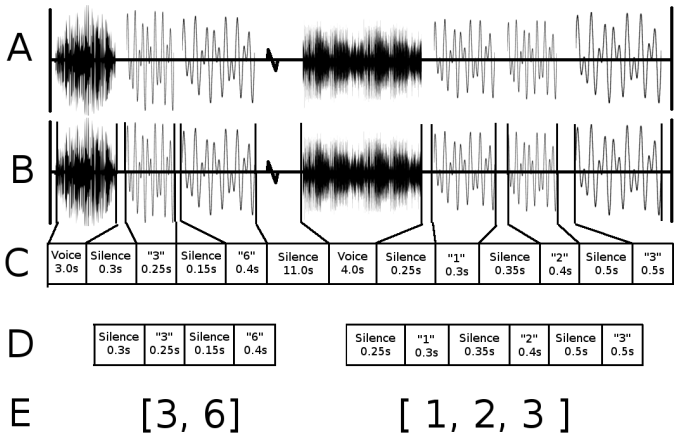

Fig. 1. CCAD overview. This figure demonstrates the flow of information throughout the CCAD process. In line A, raw input audio is shown. In line B, the audio is divided into segments such that each segment is a single type of audio. In line C, the raw audio is discarded and replaced with a tokenized representation. In line D, the tokens representing DTMF digits are separated into groups wherever a separator token appears. Finally, in line E, the dialed sequences are shown. These dialed sequences would then be compared to the patterns of known telephone numbers and recorded only if they match.

A. Stage One

Stage 1 of the method extracts DTMF information and timing information from the audio stream. Timing information is the length of any silences or voice in the audio, and is used to separate DTMF digits into meaningful groups. By default, the implementation used in this paper[1] considers voice longer than one second or silence longer than ten seconds to constitute a separator between digit groups.

Stage 1 operates by first differentiating between audio that contains DTMF signals and that which does not. Then, for audio which does not contain DTMF tones, it decides which audio is voice audio and which is silence audio.

An overview of the decision tree algorithm used in Stage 1 is provided in Fig. 2.

1) DTMF Handling: Extraction of the DTMF signals is a well understood problem. Technical documentation is available going back to the 1980s [8] describing (or containing) programs for doing so. The most popular (and simplest to implement) method is the Goertzel Algorithm [24], which can

[1]An implementation of the CCAD algorithm accompanies this paper and is available at https://github.com/spresse1/IEEE-CCAD

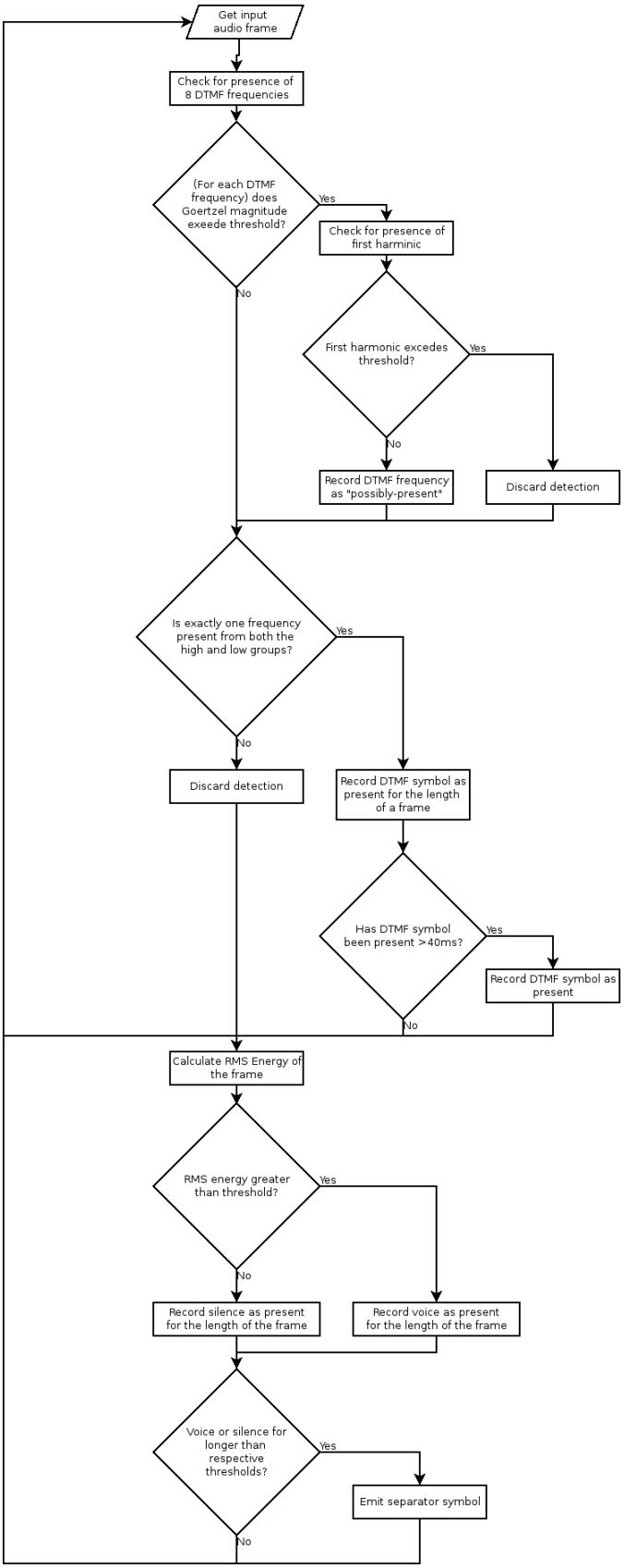

Fig. 2. CCAD Stage 1. A decision tree representation of stage one of the CCAD algorithm.

be used to determine if a specific frequency band is present. Hobbyists have implemented DTMF signal recognition as early as 1997 [11] and it is used in leading open-source software [13].

The Goertzel Algorithm is applied to the eight DTMF frequencies individually. For each frequency, the output of the Goertzel Algorithm (a unitless magnitude) is compared to a pre-set threshold. If the output is greater than the threshold, the corresponding DTMF frequency might be present. Next, for every frequency which was greater than the threshold, the first harmonic (frequency twice the original) is checked. DTMF tones are mechanically generated and will not have any output at the first harmonic. In contrast, voice or non-mechanical sounds will have a first harmonic. Therefore, if a first harmonic is detected, the DTMF frequency detection is a false positive and is ignored.

Next, the set of DTMF frequencies detected is examined to make sure that exactly two are present – one from the high set, one from the low. If more than one tone from a set is present, or a set has no tones present, the detection is a false positive and is ignored. If the frequency set passes this test, it is a potential DTMF signal.

Finally, the length of time the DTMF signal has been present is measured. ITU-T Q.23 [6] requires DTMF signals be present for a minimum of 40 milliseconds to be valid, so any shorter signals are ignored. Any potential signal which passes this test is a valid signal and is added to the signal stream.

2) Non-DTMF Handling: If a section of audio does not contain DTMF tones, we must then determine if it is silence or voice content. In order to do this, the most naive possible algorithm is used. We simply measure the Short Term Energy [14] of the section of audio. If it is above a certain volume, it is voice. If below, silence/noise. The length of each voice and noise section is tracked. If a section of audio contains voice, it is also counted towards the silence length. These lengths are both reset when DTMF signals are detected and the length of voice is reset when silence is detected. If at any point the tracked length of voice goes over one second or the tracked length of silence exceeds ten seconds, a "silence signal" is added to the signal buffer to act as a separator between sequences of DTMF signals.

B. Stage Two

In stage two of the algorithm, the DTMF signal buffer is broken into segments. A segment is any series of signals between separator signals. Each segment is then examined for validity as a possible phone number using simple pattern matching. For example, if a sequence of DTMF signals is 10 signals long (or 11 with an octothorpe as the final signal) and consists only of zero through nine, it could be a valid US telephone number and is marked as such. On the contrary, a 16-signal DTMF sequence consisting of zero through nine (optionally with the 17th signal as an octothorpe), it is not a valid telephone number – more likely a credit card number – and is ignored. As a further example, if a sequence contains A through D, star or an octothorpe (with the exception that it may end in an octothorpe), it is not a valid telephone number and is ignored.

The implementation used in this paper is written to discover domestic (US+Canada) calls only, following the North American Numbering Plan format [25], though it could easily be expanded to include the full range of international numbers defined by the ITU [26].

IV. Tests and Test Methodology

Tests were performed in two stages. In the first stage, test data was generated. Test data consists of audio and a corresponding signal stream. The signal stream represents the components that were placed into the audio and is a representation of the expected results of the algorithm when run on the corresponding audio.

Two sets of test data were generated, each consisting of 1 million audio streams. First, a set of semi-random audio streams was generated. These "type 1 tests" (or stress tests) consist of randomized sequences of DTMF signals, voice samples and noise samples. No ordering between types of audio was imposed. Voice and noise sections had a minimum length of zero and no maximum while DTMF signals were generated in lengths of 1-16, with no restrictions on signal usage or sequence. This set of input streams was used as a stress test of the DTMF detection and VAD algorithms, determining their accuracy in the worst case.

Second, a set of more restricted format audio streams was generated. The "type 2 tests" (or expected conditions) was intended to represent more typical inputs. This generated input sequences where a DTMF section was always followed by a voice or silence section exceeding the threshold for separation. This more closely models the prompt-and-response format assumed by CCAD. This set of input streams reflects the expected real-world conditions CCAD would encounter.

In the second stage of the tests, each audio stream was run through the detection algorithm implementation and the results compared to the expected results. For the type 1 tests, only the intermediate output of stage 1 of the algorithm was examined to determine success. Examining stage 2 output could have masked errors in the DTMF and VAD algorithms. For type 2 tests, only the stage 2 output was examined to determine success, as this is the real-world output of the algorithm.

Through these tests, failure was defined as any difference between the expected result and actual result. Correspondingly, a stream is successfully processed if the implementation's output matches the expected output.

V. Results

Overall, CCAD showed an excellent success rate, especially for such a naive implementation. Tests were conducted in two sets of environmental conditions. Type 1 tests ("stress tests") represent the worst-case scenario, while type 2 tests ("expected conditions") represent typical inputs.

The type 1 tests showed a success rate of 98.3 percent, while the type 2 tests showed a 99.4 percent success rate.

In order to better understand other possible improvements, we conducted an examination of the causes of failure for the first 100 failures in each test type. Failures were categorized as one or more of the algorithm having: missed a DTMF signal, missed a separator signal, added an extra DTMF signal, or added an extra separator. Additionally, failures were marked as either benign or not. A benign failure is only applicable to type 1 tests and represents a failure where the generated signal stream was different, but in which the final output (*ie*: detected envelope data) would not be different. This category only captures extra or missing separator signals adjacent to other separator signals or adjacent to the start or end of the stream. The results for the type 1 tests are shown in Table I, while those for the type 2 tests are shown in Table II.

TABLE I
CAUSES OF FAILURE IN TYPE 1 TESTS

Cause of failure	**Type 1 Test Failures**	
	Non-benign	***Benign***
Missed DTMF	0	N/A
Missed Separator	24 (24.3%)	23 (22.8%)
Extra DTMF	0	N/A
Extra Separator	40 (39.6%)	14 (13.9%)

TABLE II
CAUSES OF FAILURE IN TYPE 2 TESTS

Cause of failure	**Type 2 Test Failures**
Missed DTMF	0
Missed Separator	100 (100%)
Extra DTMF	0
Extra Separator	0

This failure analysis leads to several interesting results. First, about 37 percent of the examined type 1 failures were benign. Each observed benign failure was at the start or the end of the signal stream. Therefore, these are most likely due to differences in accounting in initial or final conditions between the test generator and the implementation than any actual error. If the ratios of failures held for the larger data set and the benign failures were corrected, the type 1 tests would have an accuracy of 98.9 percent – much more in line with the accuracy of the type 2 tests. Second, all of the type 2 failures were due to missing separators.

Finally, it is worth noting that all the failures are due to VAD issues. This indicates that the DTMF detection and the algorithm for identifying valid phone numbers is extremely robust and reliable. As the VAD algorithm (a simple threshold of the Short Term Energy of the audio) is incredibly naive, this mode of failure is expected and could easily be corrected for via the use of a more robust VAD algorithm.

VI. FUTURE WORK

The results presented here prove the viability of this approach under laboratory conditions. In order to expand this to real-world conditions, a number of improvements are required.

The single largest improvement will be the use of a robust VAD algorithm. All of the test failures were due to the selection of a very simple VAD algorithm. The use of a more robust VAD algorithm should eliminate most or all of these failures. Additionally, a more robust VAD algorithm would allow the use of real-world audio, where noise is likely to be mixed into (rather than separate from) the voice portions of the audio. It is unclear if the use of an established VAD algorithm would be suitable for this application. Most VAD algorithms are designed to mark as little audio as "voice" as possible to minimize bandwidth for audio transmission. In contrast, this application is seeking to identify what humans would understand as a contiguous section of speech.

Such a VAD algorithm may more closely parallel utterance detection. In utterance detection, like CCAD, an algorithm is attempting to find the end of a "human-understandable" section of speech, which may include short silences. However, the use of many utterance detection methods, like neural networks or speech-recognition, may be off-limits. These methods may run afoul of legal limitations, as they arguably begin to examine the content of the communication.

A second potential improvement is to use the "inter-digit time" – the length of time between tones – to infer the intent of the telephone user. For example, US phone numbers are written in groups of 3-3-4 (*eg*: 555-867-5309). People tend to dial numbers in the same format they're written [27], with longer pauses between groups of digits. This may be because they're reading them and it is simpler to remember a small part of the number, or because they dial one portion then and then look for the next part, or because they've memorized the number in this format. Other numbers of similar lengths will be broken up differently (e.g credit cards – four groups of four digits each), so the pauses between groups of digits will be placed differently.

Next, it may improve the functionality of this algorithm when dealing with values which are near the thresholds to use a Sliding Discrete Fourier Transform, rather than Goertzel's algorithm. Goertzel's algorithm, as used here, has a rather broad window (about 10ms). The use of an SDFT would allow much more precise timing of signal presence and reduce uncertainty near thresholds.

Finally, it is known that having a separate DTMF decoder in a pen register can lead to a "validity mismatch" [28], where one DTMF decoder decodes a marginal digit as valid, while the other ignores it. This can be used to circumvent or confuse pen registers. The traditional solution is to have DTMF decoding done by the telephone company and simply report the result to the pen register [29] [30] [31]. As CCAD cannot share a DTMF decoder, it cannot use this solution and a different solution is needed.

VII. CONCLUSION

In this paper we have described CCAD, the Call Contents Automatic Differentiator, an exceedingly simple and naive system for separating dialed phone numbers, which are routing information, from other data transmitted via the same signaling

mechanism (DTMF). We've shown that even such a simple implementation has a worst-case accuracy of 98.3 percent (or 98.9 percent when correcting for certain failures) and an expected accuracy of 99.4 percent.

CCAD is a new source of information for law enforcement. It operates within the existing legal framework for pen registers and provides high-quality differentiation between embedded routing information and embedded content. It therefore operates under a clear legal framework and fulfills a demonstrated need for high-quality admissible evidence by local, state, and national law enforcement.

REFERENCES

[1] *In re: Certified Question of Law*, docket no. 16-01, Foreign Intelligence Surveillance Court of Review, 2016, pp. 1-38.

[2] M. S. Gast, *T1: A Survival Guide*. Cambridge, MA: O'Reilly, 2001.

[3] L. Schenker, "Pushbutton Calling with a Two-Group Voice-Frequency Code," *The Bell System Technical Journal*, vol. 39, Jan. 1960, pp. 239-255.

[4] M. Fox, "John E. Karlin, Who Lead the Way to All-Digit Dialing, Dies at 94," *New York Times*, 8 February, 2013.

[5] International Telecommunication Union, "Frequencies to be used for in-band signaling," ITU-T Recommendation Q.22, 1988.

[6] International Telecommunication Union, "Technical Features of Push-button Telephone Sets," ITU-T Recommendation Q.23, 1988.

[7] International Telecommunication Union, "Multifrequency push-button signal reception", ITU-T Recommendation Q.24, 1988.

[8] P. Mock, "Add DTMF Generation and Decoding to DSP-mP Designs", Texas Instruments, Application Report SPRA168, 1989.

[9] C. J. Chen, "Modified Goertzel Algorithm in DTMF Detection Using the TMS320C80", , Texas Instruments, Application Report SPRA066, 1996.

[10] K. Clarkson and D. L. Jones, "Goertzel's Algorithm", 2004. [Online]. Available: http://cnx.org/contents/kw4ccwOo@5/Goertzels-Algorithm. [Accessed Dec. 22, 2016].

[11] M. Blue, "DTMF Encoding and Decoding In C", *Phrack Magazine*, no. 7, pp. 50, 1997.

[12] Zapata Computer Telephony Technology, "goertzel.c", 2001. [Online]. Available: https://sourcecodebrowser.com/zapata/1.0.1/goertzel_8c_source.html. [Accessed March 13, 2017].

[13] Digium, "dsp.c", 2002 . [Online]. Available: http://doxygen.asterisk.org/asterisk1.0/dsp_8c-source.html. [Accessed March 15, 2017].

[14] M. D. Sahidullah, and G. Saha, "Comparison of Speech Activity Detection Techniques for Speaker Recognition", 2012. [Online]. Available: https://arxiv.org/pdf/1210.0297.pdf [Accessed Nov. 29, 2016].

[15] *Ex Parte Jackson*, 96 U.S. 727, Supreme Court of the United States, 1878.

[16] *Smith v. Maryland*, 442 U.S. 735, Supreme Court of the United States, 1979.

[17] *Omnibus Crime Control and Safe Streets Act of 1968*, Title III, United States of America, 1968 (Codified as amended in 18 US Code Section 2510-2522).

[18] *Third Report and Order*, docket no. 99-230, U.S. Federal Communications Commission, 1999.

[19] *United States v. Jones*, 565 U.S. 400, Supreme Court of the United States, 2012.

[20] *Carpenter v. United States*. 585 U.S. ___, Supreme Court of the United States, 2018.

[21] Christopher A. Nolin, "Telecommunications as a Weapon in the War of Modern Organized Crime," *CommLaw Comspectus* vol. 15, pp. 239, 2007.

[22] Christopher A. Nolin, "Telecommunications as a Weapon in the War of Modern Organized Crime," *CommLaw Comspectus* vol. 15, pp. 240, 2007.

[23] Christopher A. Nolin, "Telecommunications as a Weapon in the War of Modern Organized Crime," *CommLaw Comspectus* vol. 15, pp. 242-243, 2007.

[24] G. Goertzel, "An Algorithm for the Evaluation of Finite Trigonometric Series", *The American Mathematical Monthly*, vol. 65, no. 1, Jan., pp. 34-35, 1958.

[25] North American Numbering Plan Association, "NANPA: Numbering Resources – NPA (Area) Codes," *North American Numbering Plan Association*. [Online]. Available: https://www.nationalnanpa.com. [Accessed Sept. 10, 2016].

[26] International Telecommunication Union, "The International Public Telecommunications Numbering Plan," ITU-T Recommendation E.164, 2011.

[27] D.D. Salvucci, "A Multitasking General Executive for Compound Continuous Tasks", *Cognitive Science*, vol. 29, no. 1, pp. 483, 2005.

[28] M. Sherr, E. Cronin, S. Clark and M. Blaze, "Signaling vulnerabilities in wiretapping systems," *IEEE Security & Privacy*, vol. 3, no. 6, pp. 13-25, Nov.-Dec. 2005. doi: 10.1109/MSP.2005.160

[29] "Lawfully Authorized Electronic Surveillance", J-STD-025A, American National Standards Institute, 2003.

[30] European Telecommunications Standards Institute, "Lawful Interception (LI); Handover interface for the lawful interception of telecommunications traffic", ETSI Technical Specification 101 671 v3.2.1, May 2018.

[31] European Telecommunications Standards Institute, "Lawful Interception (LI); Handover Interface and Service-Specific Details (SSD) for IP delivery; Part 1: Handover specification for IP delivery", ETSI Technical Specification 102 232-1 v 2.17.1, September 2018.