DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

VOIP SECURITY IN PUBLIC NETWORKS

by

Seylan ÇINAR

November, 2012

İZMİR

VOIP SECURITY IN PUBLIC NETWORKS

A Thesis Submitted to the

Graduate School of Natural and Applied Sciences of Dokuz Eylül University

In Partial Fulfillment of the Requirements for the Degree of Master of Science

in Electrical and Electronics Engineering

by

Seylan ÇINAR

November, 2012

İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"VOIP SECURITY IN PUBLIC NETVORKS"** completed by **SEYLAN ÇINAR** under supervision of **ASST. PROF. DR. ÖZGE ŞAHİN** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Özge ŞAHİN

Supervisor

Asst.Prof.Dr. Derya BİRANT

(Jury Member)

Yrd.Doç.Dr. Gülden KÖKTÜRK

(Jury Member)

Prof.Dr. Mustafa SABUNCU
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEGMENTS

I would like to thank to my advisor Asst. Prof Dr. Özge Şahin for her encouragements throughout this research. I also would like to thank my dear sister Gizem Burcu Çınar for supporting my research, my husband Burak Yiğit Kaya for his endless support, to my dear friends İsmail Utku Kıyak and Şenol Özkan for their editorial efforts.

SEYLAN ÇINAR

VOIP (VOICE OVER INTERNET PROTOCOL) SECURITY IN PUBLIC NETWORKS

ABSTRACT

VoIP over public networks is just a specialized internet service. In today's world, VoIP is preferred since it utilizes the existing computer networks and thus reduces costs for long-distance calls and in-company telephony network's initial setup and maintenance. As an expected result of lower costs with more features VoIP usage increases day by day.

In this thesis work, VoIP technology, infrastructure, threats to VoIP and possible precautions and solutions to these threats are researched. Among these threats, most of the technical ones have already been studied many times and certain, somewhat effective solutions to these threats are developed. Unlike many others, the weight is on non-technical attacks in this thesis and a speaker and text dependent biometric speaker verification / identification system is prototyped and successfully tested to prevent impersonation attacks.

To build the system, a collection of users sound-print files is formed from the collected samples from users and an initial voice authentication (speaker verification) before calls is aimed to be performed by means of this system. Any following work is advised to focus on integration of this system to commonly deployed production level VoIP services and a text independent version of the speaker verification system.

Keywords: Voice over IP (VoIP), VoIP attacks, non-technical VoIP attacks, speaker dependent voice verification

VOIP (VOICE OVER INTERNET PROTOCOL) SECURITY IN PUBLIC NETWORKS

ÖZ

Halka açık ağlarda internet protokolü üzerinden ses iletimi bir çeşit internet hizmetidir. Günümüzde internet protokolü üzerinden ses iletimi, var olan bilgisayar ağ altyapısını kullandığı için özellikle uluslararası görüşmelerdeki maliyeti ve şirket içi telefon hizmeti kurulum ve bakım maliyetlerini düşürdüğünden sıklıkla tercih edilmektedir. Daha düşük maliyet ve daha geniş özellikleri sayesinde beklenen bir sonuç olarak internet protokolü üzerinden ses iletiminin kullanımı günden güne artmaktadır.

Bu tezde, VoIP teknolojisi, VoIP alt yapısı, VoIP için tehdit oluşturan durumlar ve bu tehditlere karşı alınabilecek olası önlemler araştırılmıştır. Bu tehditlerin teknik olanları daha önce çok kez incelenmiş ve bunlara karşı nispeten etkili yöntemler geliştirilmiştir. Bu tezde ise üzerinde nispeten daha az durulan teknik olmayan tehditlere ağırlık verişmiş ve kimlik sahteciliğine karşı önlem oluşturması amacıyla kişi ve metin bağımlı biyometrik bir ses tanıma / konuşmacı doğrulama sistemi prototipi hazırlanarak başarıyla test edilmiştir.

Bu sistemde kullanıcılardan ses örnekleri alınarak bir veri tabanı oluşturulmuş ve her bir kullanıcının konuşmayı başlatmadan önce ses tanıma ile kimlik doğrulaması yapması amaçlanmıştır. Daha sonraki çalışmalarda bu sistemin mevcut yaygın VoIP sistemleri ile bütünleştirilmesi ve metin bağımsız türevi üzerine yoğunlaşılmasının mantıklı olacağı düşünülmektedir.

Anahtar sözcükler: IP üzerinden ses iletimi, VoIP atakları, teknik olmayan VoIP atakları, kullanıcıya bağlı ses tanıma

CONTENTS

Page

CHAPTER ONE
INTRODUCTION

1.1 Introduction to VoIP

VoIP is the most recent level of the evolvement on telephony. As the name VoIP means, it transports Voice over Internet Protocol packets. Telephony is of essential significance to the modern economy and active society. Enabling security to this service is therefore one of the most essential tasks of telecommunications. If the service is not well protected, secrecy of the transmission or availability of the service, security of the whole society is at risk.

VoIP telephony is not an entirely new idea. It is a kind of traditional telephony and it is used as a replacement of traditional telephone by its users. A replacement should provide similar level of security. Users presume that the security of VoIP telephony as granted, just like as in traditional PSTN. (Lawecki, 2007) Due to this expectation, VoIP is different from other IP services, where security is usually treated as one of the service properties configurable by the user. This makes it more important to investigate security problems of VoIP public network.

1.2 Historical Perspective

VoIP is a technology that stands for Voice over Internet Protocol and as the name would tend to suggest it also originated around the same time when the internet itself did which is around 1995. (http://EzineArticles.com/485361)

In the early days of VoIP, in order to even make a single PC to PC call both parties had to have the same sound card installed on each computer. If not then only one person could talk at one time, like say in a walkie-talkie or CB radio.

Also at first, VoIP was able to submit phone calls merely between two PCs. Then around 1998 the new call switching ability with PSTN (Public Switched Telephone Network) phones introduced by IT companies such as Cisco started the rapid growth of VoIP across the world.

As the time passed many VoIP standards started to converge and the bandwidth of internet become higher and more widely available. These developments made video telephony an expected service and even more popular than regular voice-only calls. All you need for a video VoIP call is any video supporting VoIP hookup and a webcam.

After many improvements another thing became possible with VoIP - cheap, multiparty conference and video conference calls. The days of paying astronomical amounts of money for global video conferencing become a thing of the past, again thanks to VoIP.

1.3 Literature Overview

Since VoIP is not an ever young technology and concept, many researchers studied it from a security perspective that survey vulnerabilities, possible threats and protection mechanisms that may help with these findings. D. Persky's study is a good example to the ones that explores vulnerabilities of VoIP (Persky, 2007). In other study, there is a long discussion about the threats and solutions to these threats are provided by P. Thermos and A. Takanen (Thermos, Takanen, 2008). F. Cao and S. Malik take a different approach by examining the vulnerabilities in critical infrastructure applications if they run on VoIP. In their work, the usual threats and vulnerabilities are examined, and possible ways to mitigate attacks based on these vulnerabilities are listed. The conclusion of the paper is done by providing recommendations and best practices to follow for the operators of these kinds of systems (Cao, Malik, 2006).

An operations based approach comes from D. Butcher, X. Li, and J. Guo where they overview security issues and protection mechanisms for VoIP systems with a focus on security-oriented operational practices that should be employed by VoIP providers and operators. Separating VoIP and data traffic via VLANs and similar methods, authenticated and integrity assured configuration bootstrapping of VoIP devices, securing signaling actions by means of TLS or IPsec according to the underlying protocol that is used, and the use of media encryption, SRTP are some of

the many suggestions that have been provided in their work. A brief description of how two certain commercially used systems implement such practices and guiding future research in certain directions are all part of their work (Butcher, Li, Guo, 2007).

In another paper, points out some important issues where the NIST report falls short: controversial results regarding the relative performance of encryption and hash algorithms, not using the standardized Mean Opinion Score to evaluate call quality, and the overlooked possibility of an RTP-based denial of service attack (Anwar, Yurcik, Johnson, Hafiz, Campbell, 2006). Making use of design patterns to address the problems with securely traversing of firewalls and NAT devices, detecting and successfully mitigating DoS attacks in VoIP systems, and eavesdropping protection are some of the important issues covered in this work.

Peer-to-peer usage of SIP is a challenging task which Seedorf overviews from a security perspective (Seedorf, 2006). This work lists some of the threats specific to P2P-SIP as subversion of the identity-mapping scheme (specific to the substrate overlay network), attacks aimed at network routing scheme, bootstrapping where malicious first-contact nodes exist in the network, identity enforcement (Sybil attacks), traffic analysis and privacy violation by intermediate nodes, and more such as selfish behavior of certain nodes and peers.

A multi-layer protection scheme that Reynolds and Ghosal explain, gives field researchers a way to protect their network against flood-based application and transport-layer denial of service (DoS) attacks when using VoIP. The core idea behind this scheme is using a combination of metrics from various places of the network that represent the load to continually estimate the current deviation from the network's long-term average load and successful handshakes (Reynolds, Ghosal, 2003). Similar methods have been used to detect TCP SYN flood attacks in the past with success. Evaluation of the scheme is performed via simulations that make use of various types of DoS attacks.

1.4 Aim of the Thesis

The aim of this thesis is to identify a security issue with the trending VoIP system and come up with a solution idea and a working prototype of this solution to address the security issue. To achieve this goal, one must understand the VoIP technology, its primary differences from traditional PSTN especially in terms of security issues and space for innovation. This covers going through almost all of the underlying protocols of VoIP, listing and analyzing any known security vulnerabilities and known solutions to these problems.

A detailed research shows that most of the vulnerabilities are due to the underlying IP network (Persky,2007) which are being researched on for years by computer scientists and electronics engineers whom were able to solve many technical problems such as eavesdropping by means of modern cryptography (Garg, Singh,Tsai, 2005) (IETF-RFC6189). However the long-lasting, more humane vulnerabilities, which we can refer as “social attacks” are still to be solved and pose various threats to the developing VoIP standard that are hard to be solved by pure technical approaches.

This thesis focuses on the “human side” of the security attacks after analyzing all known technical vulnerabilities and their solutions, by taking advantage of unique personal traits, biometrics, to make the underlying systems more secure. The focus is on authentication since VoIP comes with a greater mobility and passwords are known to be weak due to various reasons.(Burr, Dodson, Timothy, 2006) (Allan, 2004) (Bonneau, 2012)

The most appropriate biometric measure in VoIP is human voice since the common denominator of all telephony based services is sound, despite the possibility of transferring video or any arbitrary data. This fact also determines the core aim of the thesis as developing a user and text dependent speech recognition system, that is also known as text dependent voice authentication system (Beigi, 2011) (Furui, 2008) to ensure the authenticity of a user at the VoIP account registration step (IETF-RFC3261) using the widely adopted Hidden Markov Models (Juang, Rabiner, 2007)

that are known to produce very satisfactory results for voice recognition (Paul, 1990) (Patel, Srinivas Rao, 2010).

The focus was to cover all security issues around VoIP services including social engineering attacks which are almost independent of the technology advancements and develop a voice based authentication system, a speaker identification and verification system, prototype to prevent shoulder surfing and similar attacks on a VoIP platform.

1.5 Thesis Outline

In this thesis, an introduction to the thesis and its goals are made in Chapter 1. In Chapter 2 a brief introduction to VoIP and a comparison with PSTN is done with some emphasis on security of VoIP. Chapter 3 explains most common protocols that are used in VoIP and its applications. Known technical attacks and solutions to these attacks are covered in Chapter 4. Non-technical attacks, their fundamental differences from technical attacks are explained in Chapter 5. The proposed voice based user identity verification system and its details are explained in Chapter 6 and finally in Chapter 7, the thesis is concluded with the observations from the prototype and the gains from the research done on the topic.

CHAPTER TWO
VOIP (VOICE OVER INTERNET PROTOCOL)

2.1 What is VoIP?

VoIP is the voice and possibly video transfer over a packet switched IP network. Voice data is digitized, compressed and then split into packets to be sent over an IP network. These packets are then reassembled and used in the construction of the original analog voice data (Lawecki, 2007). All the packets that are sent over the network are significant packet, which means they contain usual data, not silence. This means efficient bandwidth usage, since the speaker actually has to say something to initiate and keep data transfer. All these packets can travel across a different path until they get to the receiving end. This is called dynamic routing. Unlike the PSTN network, there is no guaranteed and reserved bandwidth for the whole talking period (Chen, 2009). This means some packets may be lost, dropped or delivered later than their expected times during transmission.

The primary motivation behind the development of VoIP systems it its low costs. With VoIP, the cost of making a phone call across the world is much lower than the traditional PSTN way, provided that there is global network access at both points. Today, voice over IP is somewhat common and used in various scenarios;

- IP to IP
- IP to PSTN
- PSTN to IP to PSTN
- IP to PSTN to IP

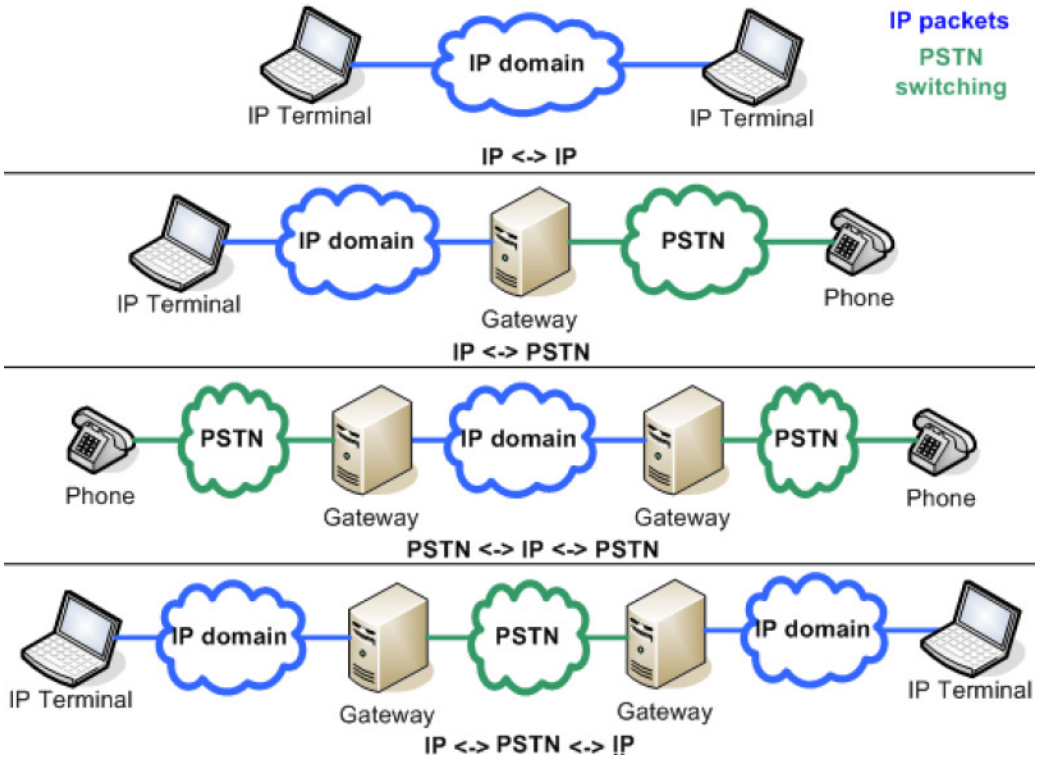

Figure 2.1 VoIP / PSTN basic scenarios (Pawel Lawecki, 2007)

Although VoIP has advantages over PSTN, it has also its own shortcomings. One of the biggest of these shortcomings is the latency. Since VoIP uses a dynamic bandwidth unlike PSTN which reserves the whole bandwidth for a single conversation and does not suffer from latency at all, VoIP communications face with latency issues from time to time due to various reasons like network slowness, server/infrastructure load etc. To address this issue, the maximum waiting time for a packet in VoIP communication is determined to be 200 milliseconds. If a packet does not reach its destination after this period, it is considered as “lost” meaning loss of voice data and sometimes interruption in communication where many packets get lost and other “new” packets cannot be received due to buffer being full and waiting for those lost packets to arrive.

2.2 VoIP vs. PSTN

When compared with each other, it is hard to determine which one is better, VoIP or PSTN. For instance the cost of maintaining and installing the infrastructure for PSTN is much higher compared to VoIP and this can be a reason to choose VoIP. On the other hand, in the event of a power outage, a PSTN network would be usable due to its standalone power supply whereas VoIP would be unusable due to lack of electricity. Despite all its shortcomings, the growing usage of VoIP across United States and Europe can be observed from Figure 2.2.

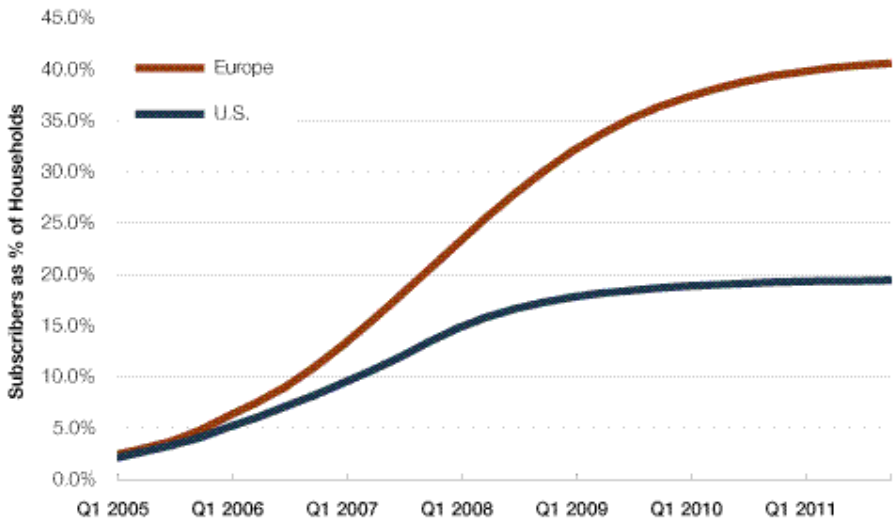

Figure 2.2 VoIP market growth in U.S and Europe (www.telegeography.com)

As seen in the figure, VoIP adoption has increased in years as its weaknesses are fixed or reduced by field's researchers and advances in technology.

To objectively compare VoIP and PSTN:

PSTN uses a circuit switched system and reserves a 64kbps bandwidth for each and every active connection. This bandwidth is reserved *and* used even if there is no actual data to transmit, when nobody speaks. This ensures a level of quality, quality of service. Though this level is fixed and cannot be changed even if there's more bandwidth available for higher quality.

VoIP uses a packet switching infrastructure. There is no such concept as a "reserved bandwidth" which means little to no data transfer when the line is silent. This means effective use of available bandwidth, which is a limited resource, and lower costs due to this reason. However, the lack of a minimal fixed bandwidth also means an unfixed level of service quality, QoS that usually varies between 4kbps and 48kbps.

Another downside for PSTN is its strong ties with the fixed proprietary infrastructure that makes it essentially non-mobile (www.iec.org). VoIP allows calls to be made from anywhere with an internet connection by means of a phone, be it a softphone or a traditional phone hardware. The new location of the user will be updated on the system. This flexibility also lowers the first setup costs of VoIP significantly compared to PSTN. While PSTN requires its own, dedicated network, meaning a not-negligible amount of up-front investment, VoIP utilizes any existing IP compatible network.

PSTN makes use of analog signals that does not have any means of compression whereas VoIP *relies* on compression at all times for efficiency.

One of the biggest advantages of PSTN comes into play on the event of a power outage. Since the network itself carries a 48V of electricity on its lines, independent from the terminals, it does not get affected by any sort of power outage at or around client terminals. This may be a big problem for VoIP on certain emergency situations since no power at the client usually means no network access, thus no VoIP calls.

PSTN assigns numbers to physical locations; client terminals whereas VoIP uses usernames, e-mail addresses, domain address etc. for identification. This means VoIP needs additional means or methods to determine the physical location of a caller, on emergency calls or similar due to its inherent nature of location independency.

Another difference is the billing between PSTN and VoIP. Since PSTN is also the network itself, physical distance plays major role when it comes to billing however for VoIP, the underlying network infrastructure, distance or location does not affect the billing process or the amount.

Another difference between VoIP and the PSTN is the internal structure. The PSTN architecture is highly centralized, complex and closed. In case of VoIP, one needs only a simple core network, as most of the functionalities are implemented in the end devices. As a consequence the VoIP network is much easier to access. However, the overall structure of the Internet and VoIP networks built on it is also very complex and covered in many RFCs. Another property of the VoIP architecture – that it is open, allows the services to be offered by different providers. In the PSTN one provider offers all the services, while in VoIP each function may be served by another subject. Different provider for access, voice services, voice mail, faxes, and data services, etc.

Development of the VoIP technology was also open and non-centralized. There was no single organization that would work on and announce some common standards. Instead of that, there were (and still are) many entities and companies and each of them came up with its own set of standards. It might be considered as a negative approach, as there is of course a lot of organizational mess. On the other hand, non-limited development enables free exchange of ideas, creative thinking and free competition of many solutions. It is just the same advantage that the open source programming has over normal programming approach.

Property	PSTN service	VoIP
switching	circuit switched – bandwidth reservation, resources are used even if there is no information to be transmitted	packet switched – no bandwidth reservation; resources are not used, if there is no information to be transmitted
services	traditional services, like phone calls, faxes, voice mailboxes, caller identification, etc.	almost all traditional services and many more – video, message, data transmission, etc.
quality	Quality of Service is guaranteed, bandwidth reservation – 64 kbps; but QoS limited to standard value and may not be enhanced	no band reservation, quality may be affected if network traffic is too high; but with a sufficient bandwidth available, quality can be better than in PSTN
mobility	originally no mobility option available, but was offered later in mobile networks	calls may be answered and originated from any place with a sufficiently fast Internet connection; user may generate calls from any place of the network, but also receive them!
infrastructure	separate telephone infrastructure necessary	shared infrastructure with data network
cost	relatively high, because of additional infrastructure and management	relatively low, as existing data networks may be used for transmission
power supply	independent power supply (48V supplied by the telephone line), telephones work even during power black out	no independent power supply, dependent on household electricity; power infrastructure needed for VoIP switches and every single desktop device
standards	well defined, common standards	no widely adopted standard, many RFCs and competitive standards covering some problems
architecture	centralized, highly complex central architecture	simple core network required, features implemented in end points, but the overall architecture complex
operators	the whole phone system is usually operated by a single company	functionalities of different OSI layers (for example Internet access and VoIP service) may be operated by different companies
compression	no compression	compression using limitations of eyes and ears to limit bandwidth; audio – silence suppression, video – motion detection
access	limited	open architecture – almost no limitations
emergency	in case of an emergency call, the caller may be localized, as each client has his/her own subscriber line	no built in emergency localization mechanism
numbering	specific to geographical location, number attached to the closest exchange and identified by the beginning of the number	uses email-similar address structure, geographically independent, identified by server's domain name or IP

Figure 2.3 PSTN services versus VoIP comparison summary (Pawel Lawecki, 2007)

2.3 VoIP Security

Although VoIP has many strong points against the traditional PSTN, it is affected by all the issues affecting its underlying IP network, internet if it is public. Major issues are usually security related. Since the VoIP traffic is just like any other packet stream, travelling across one router to another, it is vulnerable to various attacks and abuse such as worms, viruses, spam, Denial of Service attacks etc. (Thermos, Takanen, 2008).

PSTN too has its own security issues but its proprietary network infrastructure somewhat “hides” these issues to an extent and since VoIP is a more accessible, open and flexible system, it more susceptible to malicious behavior. Performing an attack to a VoIP system usually does not require anything physical or any special geographic property provided that there is network access. And on large public networks, such as internet, the attacker can easily hides his/her identity or make it very hard to find out.

Above mentioned issues become even more important when you take emergency calling into account. Even if you manage to keep the VoIP service active on a power outage, a well-aimed attack can block emergency services or prevent people from calling emergency numbers on situations requiring immediate help.

On the defense of VoIP, it should be noted that choosing PSTN for security reasons is not a real solution. PSTN networks will be using Advanced Intelligent Network (AIN) in the near future which is a system that is much more integrated to the internet (Keromytis, 2011). After this switch, almost all the underlying network issues for Internet and VoIP will also affect PSTN. Thus, security issues of VoIP are actually issues of the modern society’s telephony system and should not be disregarded in the favor of the old, inefficient telephony infrastructure.

CHAPTER THREE
VOIP ARCHITECTURE

3.1 Overall Architecture

A public network is a complex platform that consists of multiple sections. One can basically classify each entity as a client or a provider and sometimes both. A *client* consumes a service that is provided by a *provider*. A network is the infrastructure that connects all these entities together across routers, sub-networks etc. and a *public* network is a network that is accessible by basically anybody. This "public" definition can be relative such as a campus network. A campus network is essentially a private network since it is only open to campus residents but many public network concepts apply to this network since anybody in the campus has access and even some guests has access via a guest-only tunnel or similar.

3.2 An overview of VoIP

VoIP is basically the transportation of digitized sound data across an IP network by packets and reconstruction of this data on the receiving end. This process sometimes needs some specific hardware for network communication, compression (encoding/decoding) and software that conforms to specific protocols such as H.323, SIP, RTP and so on. Since all the packets are send over an existing IP network, they are subject to any inherent threats that this underlying network is subject to in addition to specific VoIP targeted attacks. The very basic VoIP definitions and protocols such as SIP and RTP does not have encryption and authenticity verification features though there are protocols developed to add these features such as SIP with TLS, SRTP, ZRTP and similar. A detailed description of these protocols can be found at Sections 3.3 and 3.4.

The intensive utilization of existing computer networks and related software gives many abilities to VoIP systems such as video conference which cannot be implemented on PSTN, at least natively and efficiently. The inherent efficiency and link independency of VoIP provides world-wide accessibility to a user via the same

number at the cost of network access just like e-mail and unlike GSM technology that powers modern world's mobile phone infrastructure which is a derivative of PSTN.

In short, although VoIP is still emerging and young, its advantages, smart and flexible utilization of existing networks, efficiency and new features make it the successor of PSTN. Since the computer networks that are used by VoIP are relatively new, it is inevitable to have issues while the network technologies grow and develop. To be able to prevent any possible issues while using a technology, one has to fully understand how it works. In the following section, the underlying protocols of VoIP will be explained thoroughly to cover this.

3.3 VoIP Session Protocols

VoIP communication requires exchanging certain data packets. Basic things such as IP discovery, domain resolution etc. uses protocols that are defined for IP networks hence they will not be covered.

Not all VoIP "protocols" are network protocols. For instance, after digitization of the analog voice signals using PCM method, this digital data is compressed using certain codecs such as G.729 by ITU-T (International Telecommunications Union – Telecommunication Standardization Sector) or similar codecs (encoders/decoders). These codecs are also part of the protocols that VoIP utilizes. Similar to protocols specific IP networks mentioned above, these codecs live outside of VoIP so they will not be covered here either.

VoIP technology is based on two kinds of protocols that serve to two basic needs of telephony: session management and voice data transmission. Consensus is to use "reliable" protocols such as TCP for session management and "low latency – low overhead" protocols for real-time voice data transmission. Since all protocols used by VoIP are application level protocols, they can theoretically utilize any network protocol.

The two most common protocols for session management are Session Initiation Protocol (SIP), developed and maintained by IETF; and H.323, which is developed and maintained by ITU-T. It is almost common sense to use these protocols over TCP with TLS. These protocols cover all session management including but not limited to call initiation, termination, user discovery and conference management.

The main protocol that is used for real-time voice and video data transmission is Real-time Transport Protocol (RTP) and its secure version, SRTP which is built on top of the existing RTP protocol. It is common to observe usage of UDP as the underlying network protocol by these protocols though the specification defines TCP as the default protocol in spite of its reliability focused design rather than the latency. SCTP and DCCP are also network protocols that are designed to be used for VoIP application but are rare to see in production environments due to various reasons.

This section will cover H.323, SIP and RTP as they are the basic and most wide-spread VoIP protocols. SRTP will be covered in a separate section since it makes heavy use of cryptographic concepts.

3.3.1 H.323

It is a protocol group of H.323 standard developed by ITU-T in order to transmit audio or image stream on a network with two or more sides, without QoS support like IP (ITU-T Recommendation, 2009). In the beginning, it was developed for multimedia conferences on local networks, and then expanded so as to include VOIP application. Various companies and institutions such as Microsoft, IBM, Intel, phone operators and ISPs attended and contributed to the definition of the standard. It stands as one of the widest and most efficient standards used for internet phone. It supports voice as well as all the other multimedia (data, video, image etc.) applications. H.323 is an umbrella standard and contains several other standards. They consist of voice coding, video coding, system control, multiplication and synchronization of multimedia streams structures. Those standards contain networks

like PSTN, Mobile, ATM, F/R, LAN, WAN and IP based internet. Some of the standards related to the systems with which IP phones have to interact are:

- H.323: is a protocol including the standards of the systems and equipment of Video Phone for LAN networks. It does not contain parameters such as QoS. ITU 96c
- H.324: is a protocol defining the standards of the system and equipment of the video phone system used in PSTN networks. H.324/M is a standard developed for cellular Mobile networks such as GSM. (ITU 96d)
- H.310: is a standard that does not contain broadband audio and video communication systems and terminals.
- H.321: defines the standards of the video phone terminals for broadband ISDN networks.
- H.322: is a standard that does not contain the video phone systems and terminals for LAN networks. It does not contain QoS parameters.

3.3.1.1 H.323 Components

The H.323 standard defines three different types of terminals.

Those terminals consist of:

- Gateway
- Gatekeeper
- Multipoint Control Unit

3.3.1.1.1 Gateway. Gateway is a set of modules working as interfaces or transition elements between PSTN and IP networks or in other words, accomplishing interworking functions. A gateway works as a "terminal" in a network providing real-time bidirectional stream between H.323-compatible terminals on a packet switching network and the other terminals on the same network or another gateway.

The other ITU terminals can be H.310 (B-ISDN), H.320 (ISDN), H.321 (ATM), H.322 (GQoS-LAN), H.324 (PSTN), H.324 (Mobile) or POTS. Gateway performs the required transformations between transmission formats (e.g. transformation between H.225.0 terminal on a H.323-compatible end and a H.221 terminal on a H.320 end) and communication procedures similar to signaling (e.g., transformation between H.245.0 terminal on a H.323-compatible end and a H.242 terminal on a H.320 end). Those transformations are defined in H.246.

Gateways bear call setup and clearing operations between IP and PSTN networks as well. The transformation between video, audio and data formats is also performed in gateways. In general, the purpose of gateways is to terminate calls between packet switching and circuit switching networks in both directions in a transparent way.

3.3.1.1.2 Gatekeeper. Gatekeeper is the network module responsible for tracking the Registration, Admission and Status of terminals and gateways with the –RAS-ETSI/TIPHON definition. Gatekeepers provide zone management and call processing/signaling functions as well.

- Address Transformations: Transformation of alias names of the network terminals into real transport names. While accomplishing these functions, the Gateway makes use of the tables that it continuously updates with the Registration messages it gets from the terminals connected to itself. These tables can also be updated by methods other than Registration messages (e.g., index services).
- Authentications: It approves or rejects the Admission Request, Confirm or Reject messages and LAN access requests of terminals. When considering LAN access

requests, call indices (call authorization), bandwidth limitations or similar criteria can be used. All the requests can be allowed to access LAN by setting this function as NULL.

- Bandwidth Management: It approves or rejects the Bandwidth Request, Confirm and Reject messages and LAN bandwidth requests of terminals.

The purpose of using Gateways is the ability to use aliases given to machines instead of machine addresses, to manage the network bandwidth and to manage network sources such as Gateway and MCU. In the original H.323 definition, Gatekeeper was designed as a unit controlling the network access during video conferences. In time, it acquired functions similar to address transformation. Bandwidth supervision appeared as a result of pricing needs.

Another service provided by Gatekeepers is the addition of security—related options to a call using various authentication methods. Q.931 or H.245 messages use in signaling can be directed by the gatekeeper so that statistical information about calls can be gathered. Phone services such as call forwarding or call transferring can be provided by Gatekeepers.

3.3.1.1.3 Multi-point Control Unit (MCU). MCUs are devices that allow more than two terminals or a Gateway in a network to join a multimedia conference. Bilateral meetings can turn into conferences and can be created via MCUs. MCU has two parts: Those are Multipoint Controller (MC, imperative to have) and Multipoint Processor (MP, not imperative to have).

MC provides negotiations on communication parameters in order to keep all terminals in call processes that will join the conference on a common communication level. MP processes media streams (mixing, switching, etc.) under the supervision of MC. MP can process a different type or a bigger number of media depending on the type of conference being carried out. With its simplest structure, MCU is composed of a single MC.

3.3.1.2 Communication of H.323 Terminals

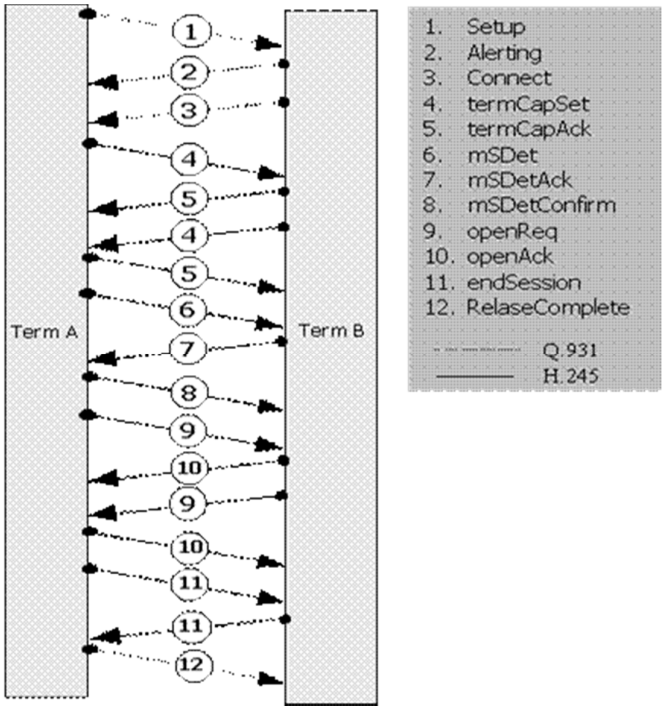

Figure 3.1Communication of H.323 Terminals

In the figure, call setup and clearing mechanism between to H.323 terminals without using a gatekeeper is explained. All Q.931 and H.245 messages that are necessary to use are listed. Each message has a sequence number assigned by the source terminal. The communication sets off by sending a *Setup* (1) message from terminal A to terminal B containing the target address. Terminal B answers with a Q.931 *Alerting* (2) message and subsequently a *Connect* (3) message in case the call was accepted.

At this point the call establishment operation is over and the H.245 negotiation operation starts. Both terminals notify the counterparty about their terminal capabilities by sending *terminalCapabilitySet* (4) messages. Media types and coding methods can be given as examples of terminal capabilities. Terminals answer these messages with *termCapabilitySetAck* (5) messages. During the session, terminal capabilities can be re-sent at any moment.

After this step comes the determination of *Master/Slave* (6-8). Both of the H.245 Master/Slave determination procedures are used for eliminating conflicts between terminals being able to serve a conference as MC or trying to open a bidirectional communication channel. In order to determine the master and the slave terminal, both terminals transmit random numbers to each other via H.245 *masterSlave* determination messages. All H.323 terminals are supposed to be able to operate as both master and slave.

After the Master/Slave determination procedure, both channels send messages to each other (9-10) in order to open up a logical channel. While audio and video channels open in one direction, data channels are bi-directional. Terminals are free to open as many channels as needed. The flow in the figure represents a single channel. The procedure applies to all channels to open.

The closing of the session (or communication) sets off after the *endSession* message to be sent by one of the parties.

3.3.2 SIP (Session Initiation Protocol)

SIP is the abbreviation for session initiation protocol and it is used to manage communication session with two or more participants (RFC 3261, 2002). It is an application layer protocol developed by IETF and works independent from the underlying transport protocol such as TCP. Known implementations mostly use TCP, UDP and SCTP in descending order. It inherits its many properties from HTTP, the protocol that powers the modern web, such as being text based, usage of URIs, request/response architecture with header and body sections including certain standardized headers and status codes. The latest version of SRTP as of today is RFC 3261 and it is a permanent element of the architecture used for IP-based multimedia streaming services in cellular networks since November 2000. A reference implementation is provided by the US National Institute of Standards and Technology (NIST) in Java programming language.

The URI protocol identifier for SIP and secure SIP are “sip:” and “sips:” and they use the ports 5060 and 5061, respectively. SIP is used to orchestrate communication sessions including session initiation, termination, port changes, inviting new participants to a session and so on. It is also used for event subscription and notifications especially in instant messaging applications and similar.

The protocol focuses on initiating and terminating calls. The rest of the PSTN features and other new features can easily be implemented with the help of specific proxy servers and user agents (soft phones) thanks to the flexible nature of the protocol. It piggybacks on RTP for real-time data transfer duties and utilizes session definition protocol (SDP) in its message body to define the properties of the real-time stream.

Although the protocol is a decentralized peer-to-peer protocol it makes use of certain centralized elements for the sake of consistency and discoverability reasons. A typical SIP implementation has the following elements:

3.3.2.1 User Agent

A user agent in an SIP context is the most used and the only essential unit. It is an end-point which receives and creates SIP messages. It corresponds to the traditional telephony unit that is connected to the land line in houses and probably is the only user facing element of an SIP chain.

A user agent acts both as a client (UAC) and the server (UAS) according to the context and the call status. This behavior requires a non-trivial logic so even if some implementations come in the form of hardware that just looks like a traditional phone, the real job is done via firmware which simply is certain software.

Just like in the HTTP a user agent populates the “user-agent” header field automatically with its predefined, unique user agent string which gives any other peers some information about the type and capabilities of the agent.

3.3.2.2 Proxy Server

A proxy server acts as a middle man, mimicking the dynamic UAC and UAS behavior of user agents with following or enforcing certain logic such as routing the requests to the closes client/neighbor, preventing calls that are not allowed or changing/adding certain header fields for a better service.

3.3.2.3 Registrar

A server acts similar to a phonebook. It processes the REGISTER requests and binds users and their IP addresses to certain URIs. These servers make this data available to proxy servers or redirect servers for use.

3.3.2.4 Redirect Server

A server that responds to their clients with only 3xx, redirect responses. This kind of servers can be used for network load management and to redirect users to other domains for external endpoints.

3.3.2.5 Session Border Controller

A controller that sits between user agents and any other SIP related entity to enforce certain rules (control) for the sake of security, network isolation and similar features.

3.3.2.6 Gateway

As the name implies, these entities interface a certain VoIP network to other networks such as an external WAN or PSTN, a different kind of network.

SIP follows all low-level specifications of the inherited HTTP protocol such as line ending policies, status information in the first line etc. A major divergence from the HTTP protocol is the 1xx provisional status codes that indicate the message is received, necessary action has been taken but not completed yet and a final response should be waited for. This is similar to HTTP 102 response though more involved

since it requires the sender to wait for a final answer from the same party and keep state which makes it more susceptible to memory based denial of service attacks.

Another divergence from HTTP is the ACK and PRACK verbs/messages that are used to acknowledge that the message is received reliably. This is necessary since the server and client roles in SIP are intermixed unlike HTTP which has a strict definition of a server and a client and is built upon a fire-and-forget philosophy.

SIP verbs or commands that are defined as the writing of this thesis are as follows:

- **INVITE:** Invites a client to participate in a call. Used to start calls.
- **ACK:** Acknowledges that the final response for a request has been received successfully.
- **BYE:** Terminates an ongoing call. Can be sent by any participant.
- **CANCEL:** Cancels pending requests.
- **OPTIONS:** Gets the capabilities of server.
- **REGISTER:** Registers the address provided in the “to” header to a server and with the provided URI.
- **PRACK:** ACK for 1xx responses.
- **SUBSCRIBE:** Subscribes to certain event notifier.
- **NOTIFY:** Notifies the subscribers about an event.
- **PUBLISH:** Publishes an event to the server such as a presence change.
- **INFO:** Sends information about an ongoing session which does not alter the state or the session.
- **REFER:** Used for transferring calls.
- **MESSAGE:** Used to send messages for applications like IM over SIP.
- **UPDATE:** Updates the state of the session with a new information.

SIP response codes are grouped as follows:

- **1xx Provisional:** Request has been received and being processed. Keep waiting for a final response.

- **2xx Success:** Request has been received accepted and successfully processed.
- **3xx Redirection:** Further action (commonly another request to another URI) should be taken by the client (usually).
- **4xx Client Error:** Request is not valid or cannot be understood by the server so it will not and cannot be processed.
- **5xx Server Error:** Server had a permanent or temporary issue that prevented it from fulfilling an apparently valid request.
- **6xx Global Failure:** A new response type that is user centric and used to handle a declined call, globally non-existing user or a globally busy user.

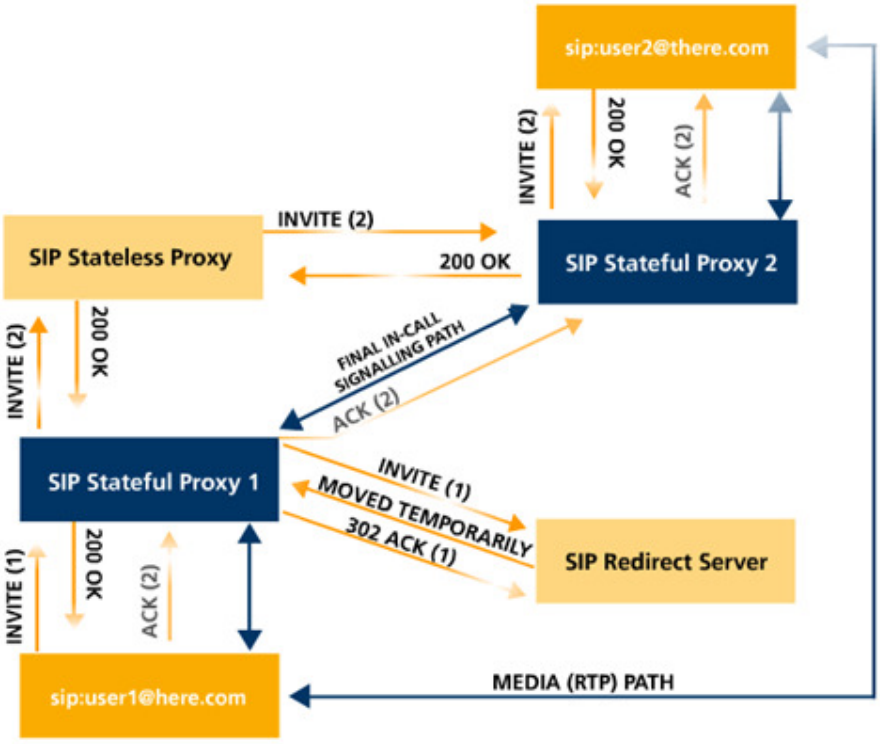

Figure 3.2 VoIP market growth in U.S and Europa - http://en.wikipedia.org/wiki/File:SIP_signaling.png

3.4 VoIP Data Transmission Protocols

In the data transmission phase, there are mainly three protocols. Those are RSVP (Resource Separation Protocol) used for separating resources, RTP (Real-Time Protocol) used for real-time data flow and RTCP (RTP Control Protocol) assuring the control of the RTP.

Before transmission of data in the system, signaling with SIP or H.323 takes place. Then, a part of the system resources get allocated for VoIP meeting by RSVP. After that, SDP and the terminals are notified about which UDP ports to use in order to use RTP and RTCP.

3.4.1 Resource Reservation Protocol (RSVP)

As the name suggests, RSVP is used to allocate resources needed for opening a session on Internet. Since it is a protocol without IP connection, road establishment does not occur. Therefore no specific band width is allocated for these roads. RSVP is designed to provide the band width required for the stream on the established roads. Even though RSVP does not get involved into redirecting activities, it uses several versions of IP as a carrying mechanism as with ICMP and IGMP. RSVP runs resource separation protocols for a multicast group.

3.4.1.1 Working Modes of RSVP

RSVP has two working modes. These are road establishment mode and reservation mode.

3.4.1.1.1 Road Establishment Mode. In this mode, RSVP runs either one of unicast and multicast working procedures. As explained above, resource relocation procedures are run. It needs the service quality requests of the parts receiving RSVP stream for flow. The application running on the receiving side decides which QoS profile will be transmitted to RSVP. After the receipt of the request message, RSVP sends requests messages to all nodes along the data flow. RSVP is also used for

transmitting QoS request messages issued by redirectors to nodes and for the relocation of the necessary resources for these request messages at each node.

3.4.1.1.2 Allocation Mode. In this mode, the receiving party informs the sending party and the intermediate elements (such as redirectors) its own QoS requirements. This mode is also known as the reservation mode.

3.4.2 Real-Time Transfer Protocol (RTP)

RTP is a protocol developed for fulfilling network carrying functions like the carrying of the audio and video data from one end to the other in real time. RTP works upon UDP. It uses multiplexing and heading control mechanisms of UDP. In spite of this, RTP can work with other low level protocols.

Another important feature of RTP is to carry out the data transfer of multiple users in multicast environment. This way, audio and video conference applications are possible to authenticate.

RTP facilitates the synthesis of the audio or video on the data receiving side thanks to its serial numbers. Besides synchronization operations can be easily performed by the time stamp (tag) included in the RTP.

RTP naturally allows the use of translators and mixers. Translators perform the transformation and coding of the data (or payload) that is transmitted into another format and coding. Let's assume there is a system generating video at 1 Mbps. The data generated in this system can be transmitted appropriately and simultaneously on a 128 Kbps data bus by an RTP translator. The RTP translator permits the interaction of the 3 stations seen above. Also the data coming from these stations is packed in accordance with the band expansion constraints of the system.

On the other hand, RTP mixers allow the transformation of data coming from multiple sources into a sing data stream. Especially mixers that attend audio operations do not reduce the signal quality reaching the receiver. They only combine several signals into one in accordance with a certain format.

3.4.3 Real Time Control Protocol (RTCP)

Following the reservation operation by RSVP, data packages start to flow between terminals. Then the RTCP steps in and allow the terminals to notice the level of service quality they can provide and receive. RTCP is a protocol working in association with RTP. It is used for feedback about the data transmission quality.

CHAPTER FOUR
VOIP RELATED TECHNICAL ATTACKS

VoIP is subject to many security threats some are inherited from its IP network base and some are specific to VoIP itself. VoIP specific threats and attack types can be grouped as follows:

- Denial of Service (DoS)
- Eavesdropping
- Spoofing
- Replay Attack
- Man in the middle attack
- Call Hijacking
- Call redirection

4.1 Denial of Service (DoS)

A DoS attack is an attack that aims to prevent a service from running, or preventing clients from accessing and/or using a service. This is done by depleting various resources such as bandwidth, disk space, CPU or memory. Another way to achieve this goal is to disrupt network access via physical means or changing network configuration in a malicious manner.

A DoS attack that is targeting the bandwidth is one of the hardest attacks to mitigate. The best possible way is to prevent the attacker from sending or injecting packets into the network at all which is usually not possible, especially on a public network. A more feasible way is to detect excessive packets from a certain IP address and stop further processing of these packets. This IP address based protection fails against distributed denial of service attacks or DoS attacks that also make use of IP spoofing or forging.

4.1.1 Types of Denial of Service Attacks

4.1.1.1 Buffer Overflow Attacks

This is one of the most wide-spread DoS attacks. A buffer stores a certain type of data in memory. A buffer overflow attack sends a large data to the server that the programmer did not expect nor protected itself for causing a portion of the memory to be overwritten by the received payload. This may lead to arbitrary code execution in the worst case and buffer full errors causing new data to be dropped on the best case.

4.1.1.2 SYN Attacks

This type of attack is again one of the most wide-spread DoS attacks and makes used of the naïve behavior of servers. In its TCP version, an attacker sends lots of "SYN" packets to the server from various, fake IP addresses causing the server to reserve a connection state for each packet leading to memory and/or CPU depletion and in some cases bandwidth depletion too. In its VoIP version, the most common scenario is to flood an SIP UAS or proxy server with lots of INVITE messages, usually using fictitious, non-existing user info. These attacks have similar consequences with their IP based versions.

To secure the system from a DoS attack, the first thing to do is to analyze the attack and pinpoint the source of the attack if possible. DoS attacks can originate both from the internal or external network. An external attack can easily mitigated by cutting all external access from the internal network though this will isolate the service from the outer world, which is not desirable in most cases. An internal attack is harder to mitigate since it has a much faster access to all system resources and cutting access for an unknown internal attacker might be harder. In spite of these problems, an internal attack is much easier to pinpoint since the attacker should use a known entry point on the internal network.

4.2 Eavesdropping

Eavesdropping is the involvement of an unwanted third party into a conversation between two or more people. This was a very common issue on PSTN and is also an issue for VoIP. People exchange private information such as identification numbers, credit card numbers, or personal details over phone calls and this clearly shows how serious can an eavesdropping attack be.

In today's world, network stalking tools, such as Wireshark are easily accessible via the internet and they are even free to use with tutorials and have built-in features for data extraction from known protocols such as RTP. This makes performing the attack much easier with a much lower barrier for trial.

This attack requires access to the network but not necessarily a physical access. For instance usage of a hub device instead of a switch makes any packet sent to a specific device connected to that hub available to all other devices that are connected to the same hub. Another example would be unsecured wireless networks where even a passerby can gather private information without much hassle. Of course any ability to tap into the physical connection would allow this attack to be performed too. Also, any kind of man in the middle attack allows the attacker to eavesdrop calls.

Unlike PSTN, an eavesdrop attack against VoIP can have a wider range of impact even though the entry point to network does not have direct access to central resources. This imposes a greater security risk than its predecessor, PSTN. The main issue is the lack of default encryption and authentication mechanisms in the RTP protocol. For a secure communication, SRTP should be enforced, where it is still not enforced. Due to this reason, any person who has network access can record and listen to any VoIP calls that uses RTP for media transport, via Wireshark or similar application that allows packet extraction. This lack of authentication and encryption also paves the way for another attack, replay attack, which will be covered in following sections.

4.3 Spoofing

Spoofing is the delivery of packages with a wrong IP or MAC source address. The attacking person can hide his own IP address or point another person as performer of the attack. Aside from letting the attacker seem to be a reliable user, this allows attacks such as listening to the network or capturing it.

The biggest risk in a spoofing attack is ID theft. For example, a client may give a phone call to a place and give her/his credit card number in order to place an order. However, (s)he may possibly be sharing her/his credit card and ID information with the attacker because of a spoofing attack. An attack of the sort stands as an example of "man-in-the-middle" type attack.

Another type of spoofing is performed by modifying the caller ID or CLID (call line identification). It is realized by changing the phone number of a fictitious or real user. And that can cause the systems verifying according to the calling number, judge wrong.

An example of spoofing attack is shown below. Users A and B talk through a VoIP system. User C causes the termination of the call by sending a "BYE" message to the user B on behalf of the user A. While the user B believes the call was ended by the user A, the user A is not aware of the reason why the meeting was over.

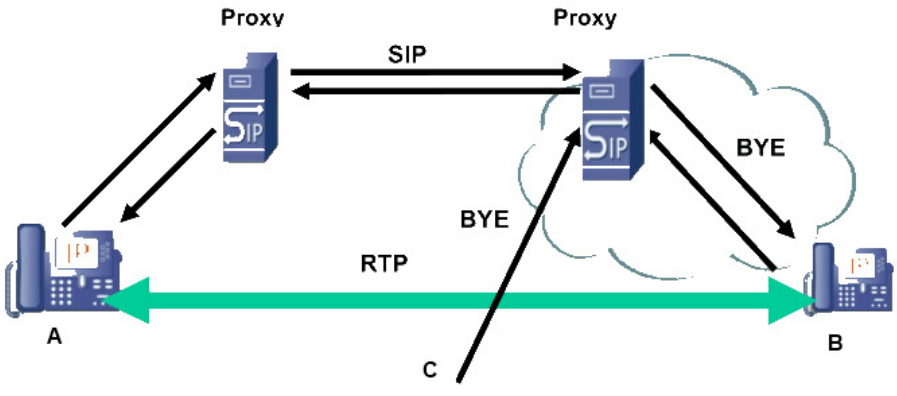

Figure 4.1 An example of spoofing attack

4.4 Replay Attack

It is a type of attack realized by repetitive transmission of the same audio package. A person intervening a call between two people records some or all of the talk and then performs the attack by transmitting to the receiver party the packages (s)he received. The riskiest part of such an attack is that the talker may be sharing her/his personal information or approving an important operation. The attacker can cause the approval of undesired operations by transmitting to the receiving party repetitively the audio packages containing the confirmation sound of the talker.

4.5 Man-in-the-middle Attack and Call Hijack

The man-in-the-middle attack means the ability of the attacker to read and modify the messages of the talkers without their notice. It can as well be used to realize other types of attack such as DoS or wiretapping.

On the other hand, Call Hijack is the case where the attacker can substitute on of the talkers. A person performing a Call Hijack attack can transform it into a man-in-the-middle attack in order not to raise any suspicion.

These attacks can take place in different ways:

- Registration manipulation
- Call direct

4.5.1 Registration manipulation

In this type of attack, the attacker redirects all incoming requests by changing the user records in the systems. By changing the “from” title of a SIP request, a fake records can be easily generated. Since the UDP protocol used for these registration requests is a connectionless one, it can easily be spoofed.

SIP Registrar server does not have to verify the User Agents requesting for registration. Anyway, in general, there is no verification in intranet which is considered to be secure.

An example is shown above about an attack regarding the modification of the registration records. The attacker first renders the Bob user's terminal unusable by performing a DoS attack to the user. (S)he the requests to register to SIP Registrar server by taking the Bob user's username. Thus (s)he changes the registration information of the Bob user. After this step, all requests arriving at SIP proxy where Bob is registered will be directed to the attacker's IP. When Alice wants to call Bob, an "INVITE" message is transmitted to proxy so as to call Bob, as can be seen in the figure. Proxy then redirects this message to the attacker's IP that it thinks is Bob's. When the call is established, Alice gets in fact in talk with the attacker while thinking that she does with Bob.

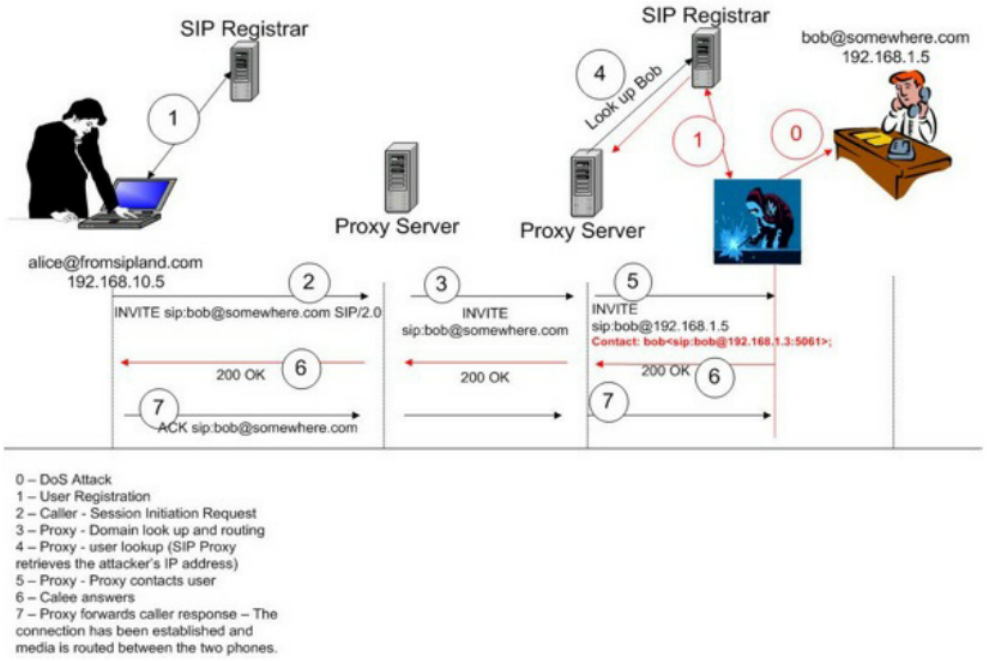

Figure 4.2 An example for Registration manipulation

4.5.2 Call Redirect

SIP 3XX answer codes are used for redirection. It informs the requestor on what should be done in order to execute the request and redirects to the related place. 3XX answer codes in SIP attacks are used for fake answers.

For example, a user that is registered in SIP system makes an "INVITE" request. The attacker sends "3xx" as answer to the user that made the request. Here the attacker has replaced the User Agent or one of the SIP components. The communication of the user receiving the "3XX" message is the redirected to the party designated by the attacker.

Examples of 3XX messages are "301-Moved Permanently", "302- Moved Temporarily", "305- Use Proxy".

A typical example of this type of attack is the call hijack executed by using a 301 message. In this scenario, as soon as the attacker notices the request of the user by an "INVITE", (s)he sends a 301- moved permanently message to the user and thereby redirects the request to herself/himself. Thereon the user establishes a connection with attacker in order to realize the SIP request.

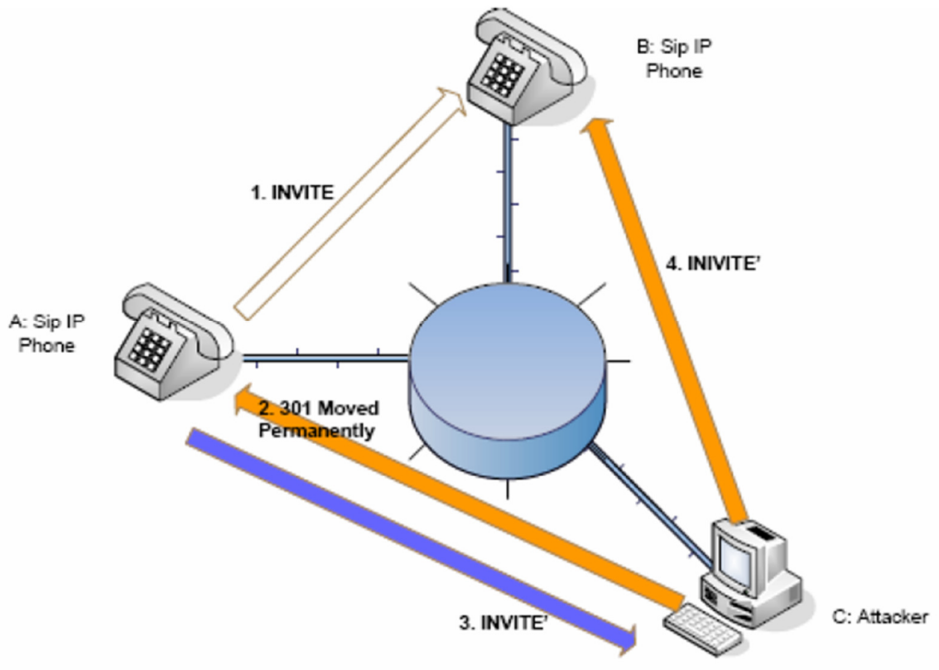

Figure 4.3 A simulation for call forwarding attack by using 3xx answer codes

A call forwarding attack by using 3xx answer codes is simulated in figure 4.3.

Call hijacking attack can also be realized by the use of the SIP 302 answer code. The attacker again sends a 302- moved temporarily message upon a SIP request and indicates that the user is temporarily redirected to her/his own IP. Therefore (s)he ensures that the user temporarily establishes a connection with her(him) for realizing the SIP request.

Another example can be achieved with the use of "305 – Use Proxy" message. In this type of attack, the attacker substitutes the proxy. When a user submits a SIP request to the proxy (s)he is registered at, the attacker tells, by sending a 305 message to her/him to use the proxy at her/his own IP for the request Therefore the user sends further requests to the attacker instead of the proxy (s)he is registered at.

4.6 Spam over Internet Telephony (SPIT)

Bhan, Clark, Cuneo, and Ramirez (2006), define Spam over Internet Telephony as below:

"Analogous to the email spam problem in data networks, security analysts have envisioned a major attack of voice and video messages in VoIP networks. Even though mass advertising attacks have been launched by advertising agencies on the regular PSTN network, the complexity and costs of doing so are prohibitive for mass harassment. However, SPIT becomes a major issue without traditional telephony lines. The access to millions of internet phones and traditional PSTN phones via the internet at extremely low costs is a resource just waiting to be abused by attackers once penetration of VoIP services have gained significant momentum. SPIT poses a potentially critical threat to VoIP services as millions of unwanted voice messages (i.e. advertisements) could overwhelm customers.

Although this attack seems extremely similar to email spamming attacks, and there are advanced solutions such as blacklists and quarantines developed to combat

email spam, applying those technologies to VoIP networks would be extremely hard given its real-time nature and difficulty in deciphering the content of the message. SPIT attacks that target the PSTNs from the VoIP networks would almost be impossible to block. There are also concerns of session hijacking in VoIP, whereby an attacker would be able to capture a video conference channel and transmit advertisements instead. Similar attacks would also be possible on voice conversations which could be hijacked for impersonation or broadcasting mass messages."

4.7 Solution Proposals

4.7.1 General Precautions

All necessary precautions for the security of an IP network are also necessary for the security of a VoIP network. In a standard VoIP network, voice, multimedia and data packets reside in the same network, thus affect each other. For instance, an attack aimed at the IP network which carries data, will also harm all VoIP traffic on the same network. Thus, the very basic and preliminary precaution against this situation is to isolate data and voice traffic from each other. To achieve this, separate VLAN's should be used for the standard data traffic and the VoIP traffic. With the aid of ACLs, any intercommunication between the data and voice network should be kept to a minimum. VLANs are virtual networks, so there will be no physical separation though a separation at software layer is more than enough and necessary to keep the common network advantages of VoIP.

Another solution is to restrict all kinds of access (telnet, SSH etc.) to the VoIP servers on the network from unauthorized people. Some of these restrictions are change of port numbers, restriction to certain IP addresses and similar but not limited to these.

Another essential solution is to use a firewall. A firewall protects resources residing on local area networks from possible attacks that originate from other networks. It basically manages the network traffic between internal and external

networks based on certain set of rules and can be thought as a restrictive gateway. A firewall can either be software or a completely separate hardware. They drop or forward packets coming through and going out from a local network and their main purpose is to restrict access which generally makes them a whitelisting solution instead of a blacklisting solution. Despite this generalization, they can be used either way according to security needs of the network.

4.7.2 IPSec

IPSec is used to secure network traffic at network layer. It is a collection of protocols developed by IETF (RFC 2401). It can be used to secure any kind of application layer protocol, regardless of the transport protocol they depend on. So it can be used to secure RTP traffic as well as SIP traffic over TCP or UDP. IPSec provides authentication, integrity verification, protection against reply attacks and eavesdropping attacks.

IPSec satisfies the following security needs:

- **Data Confidentiality:** IPSec ensures data confidentiality using encryption. Data encryption prevents third parties from reading/understanding the data. DES, 3DES and AES are the algorithms that IPSec makes use of.
- **Data Integrity:** IPSec ensures that the data reaches its destination without any modifications or manipulations. This is performed by means of secure hash algorithms. A hash digest of the message is generated and verified on both ends to ensure this property. IPSec makes use of MD5 and SHA-1 algorithms.
- **Message Origin Authentication:** IPSec has an authentication header which both serves for data integrity and origin authentication purposes by means of HMAC-SHA-1 method. This ensures that the origin is the one who claims it to be and the data is intact.

Core IPSec protocols are: IKE (Internet Key Exchange), ESP (Encapsulating Security Payload) and AH (Authentication Header). IPSec establishes a secure tunnel between two points using these protocols. Initiating party first defines what kind of

packet traffic will be protected and will be transferred via this tunnel. Right after this, the parameters that define the characteristics of the tunnel are set. If a party will send traffic that is predetermined, it will use a tunnel whose properties are predefined and utilizes this.

The protocols and algorithms that will be used for the tunnel are determined by SA (Security Associations). The three core protocols that IPSec utilizes to secure the traffic are explained below:

- **IKE (Internet Key Exchange):** It is responsible for the sharing of security parameters and encryption keys. IPSec uses symmetric algorithms for data encryption. These algorithms are very fast for streaming large amounts of data and are pretty secure once a secure key exchange is performed. IKE serves for this secure key exchange operation.
- **AH (Authentication Header):** AH, provides data integrity and authentication features at the same time. This data is embedded inside the subject data. After the ESP protocol, AH has lost its important pretty much.
- **ESP (Encapsulating Security Payload):** ESP ensures security by encryption and authentication. Many IPSec applications utilize ESP. ESP can also provide personal privacy with encryption.

Using IPSec to secure SIP messaging brings additional load to the packet header.

4.7.3 Transport Layer Security (TLS)

TLS is used to secure SIP messages at transport level and requires a TCP connection. TLS is defined in RFC 2246 and plays a very important role in establishing secure sessions. When using SIP over TCP, TLS or SSL can be used to secure the communication between servers and clients.

TLS has two parts:

- TLS record protocol - uses symmetric key encryption

- TLS handshake protocol – authentication between the client and the server, key and encryption algorithm negotiation

If the scheme in the SIP URI is SIPS instead of SIPS, it means TLS should be used just like the case for HTTP and HTTPS. If the requested party does not support TLS, it refuses the connection. TLS requires a reliable transport layer thus it cannot be used with UDP and is usually used with TCP. This makes it impossible to be used on an UDP based SIP communication.

4.7.4 Secure Real-Time Transport Protocol (SRTP)

SRTP is a protocol specifically designed for VoIP for media packets to be transferred securely.

SRTP is used to secure RTP packets and has been published on March 2004 by IETF via RFC 3711. It aims to secure RTP which is used to transfer real time data such as audio and video streams. It adds confidentiality, authentication, protection against reply attacks and similar security measures to the existing RTP protocol by modifying or wrapping those packets.

SRTP provides a high throughput and keeps the additional load on the existing RTP packets while providing these security features under a single, convenient stack. SRTP is independent from the underlying algorithms used for encryption, hashing and key exchange. It can work with any algorithm, though Multimedia Internet Keying (MIKEY) is specifically designed to work with SRTP and for this reason, it is the leading protocol that is used for key management in SRTP implementations and applications.

All features in SRTP such as authentication and even encryption are optional. If encryption is used, the default algorithm for it is AES-CM (Advanced Encryption Standard – Counter Mode) with a 128-bit key length.

The default algorithm for authentication in SRTP is HMAC-SHA1. Authentication key length and authentication tag length are 160-bits and 80-bits respectively.

SRTP has many advantages over RTP which can be summarized as below:

- Security for RTP and RTCP via payload encryption
- Ensuring message integrity with reply attack protection for RTP and RTCP
- Reduction in the amount of cypher using the same key via periodical key renewal
- A secure session key generator via a cryptographically secure pseudo-random number generator on each end
- Security for unicast and multicast RTP applications

A typical SRTP header is shown in Figure 4.4:

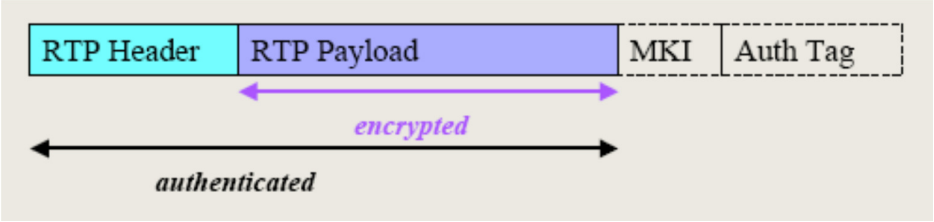

Figure 4.4 SRTP Header

MKI (master key identifier), defines which key will be used. Before moving onto the flow of SRTP, a brief look at the cryptographic parameters that it makes use of would be valuable:

- SEQ: Sequence number (16 bit)
- ROC: Roll Over Count (32 bit) – Number of times that the RTP sequence number is reset
- Index (48 bit)
- i=216 * ROC * SEQ
- Replay list
- Master key
- Authentication algorithm identifier

- Encryption algorithm identifier

4.7.4.1 SRTP Flow

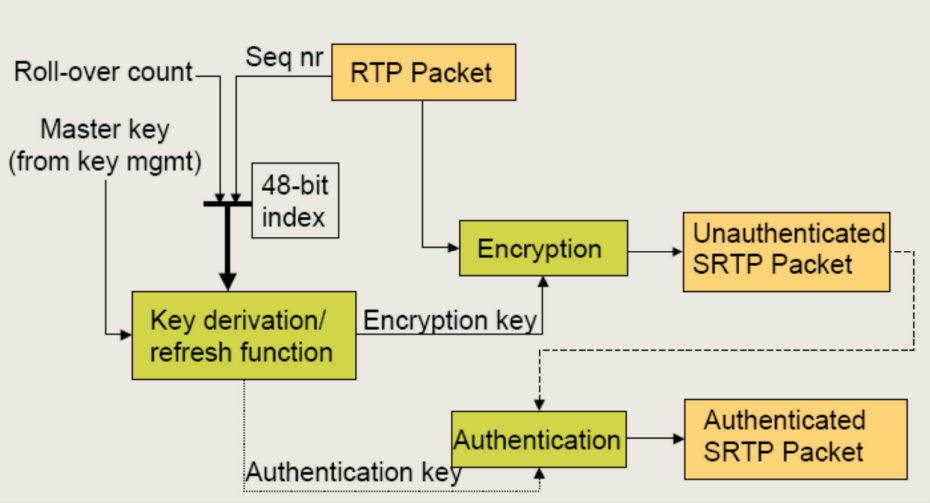

Figure 4.5 SRTP Flow

As seen in the Figure 4.5, the 48-bit index is calculated using RTP sequence number and ROC at first. Index and master key are processed through a key generation and renewal function to obtain two new keys: encryption key and authentication key. Encryption key is used to encrypt RTP payload and the resultant cypher is combined with the authentication key to obtain the final authenticated SRTP packet.

4.7.4.1.1 SRTP Encryption. AES-CM is used by default for RTP packet message encryption. Encryption system combines the SRTP packet index and the 128-bit key in a pseudo-random bit sequence. Each bit sequence is used to encrypt a single RTP packet. Encryption of an RTP packet consists of two separate steps:

1. Generation of the key stream corresponding to the packet
2. Bit-by-bit XOR operation of the key stream and the RTP payload to obtain the final encrypted SRTP cypher

4.7.4.1.1 SRTP Authentication. Calculation of the authentication tag can be defined as follows: sender calculates the authentication tag and adds this to the end of the message. Receiver uses the previously selected algorithm and the session authentication key to calculate a new authentication tag and compares these with the received message and authentication tag. If all matches with each other, the message is claimed to be authenticated and if not, authentication fails. HMAC-SHA1 is used in this process as default.

4.7.4.1.2 SRTP Key Management. An authentication and encryption key are required to use SRTP to its fullest extent. At first, these keys are derived from the master key and the key renewal algorithm renews these keys in r^{th} packer intervals. This key generation algorithm is based on AES-CM by default.

CHAPTER FIVE
SOCIAL ENGINEERING

Social engineering can be defined as deceiving people to give out their private information or provide access to a restricted system and is a serious threat to many "secure" networks. Although there is more than enough information on social engineering, means of protection are usually far from sufficient. Influencing, forcing, developing delusive relationships or using methods to decrease loyalty, honesty, ethical values or responsibility are all methods that are used for the success of social engineering attacks and to prevent these strong security policies, training and good incident response methods are necessary (Oosterloo, 2008).

5.1 Definitions

In computer security terms, Social Engineering is the combination of methods that exploit gaps and vulnerabilities in relationships and human behavior to bypass security measures. The relationship, in this context can be any of the following: inter-personal, between a person and an organization or between two or more organizations (Anderson, 2000). The gaps in human behavior can be explained as daily, and mostly unconscious, actions that are separate from conscious intentions that may present vulnerabilities in terms of security.

Typical targets for a social engineering attack against an organization are the personnel that the attacker can abuse or manipulate. Victim profiles can be summarized as follows:

1. **Directly accessible personnel (Service personnel, employees that answer phone calls etc.):** Employees that represent the organization to the outside world and are in contact with customers and providers due to the nature of their assigned jobs.
2. **Important personnel (Administrators, personnel that have access to private information):** Employees that have non-trivial permissions or have access to sensitive information due to their job requirements or similar reasons.

3. **Personnel having sympathy:** Employees that may use their reputation in the organization or exceed their authority to help and support customers.
4. **End users in need of support:** Users that have access to the organizations systems due to the service they receive but may not be able to differentiate an impersonating attacker from a legit organization support personnel due to lack of knowledge.
5. **Tricked, deceived or persuaded personnel:** Personnel that still work for the organization but are less loyal to the organization or its employees by external means.

Attacker profile, on the other hand, may vary according to the target and the method. Some of the most common methods are as follows:

1. **Authoritative approach:** Persuading that he/she is authorized, a high level manager, or a privileged customer.
2. **Offering help:** Tricking customers or employees in need that he/she is the authorized personnel on the issue.
3. **Finding similarities or common points:** Creating virtual social connections (relative, same occupation, common friend, same community etc.) between him/her and the employee.
4. **Offering a return:** Offering something in return for a favor.
5. **Abusing loyalty and honesty:** Convincing an employee that the organization will be harmed if he or she does not do what the attacker says.
6. **Making use of lack of loyalty:** Luring employees with low loyalty by means of persuasion, deception or tricking.

5.2 Methods

The definitive property of social engineering attacks is keeping the illusion that the attacker's actions are legal and legit. For this reason, the methods can differ in terms of their nature or content according to specific conditions of the environment. In this section however, a standardized classification will be made.

5.2.1 Making up fake scenarios

This type of attack happens usually over phone conversations. The attacker forms a fake scenario to read the necessary private information (such as personal identifying information, passwords, security policies etc. to be used in the next step) for his/her own purposes between the lines of the conversation going on according to the made up scenario. Since authentication via phone are done using information that is available over other channels (such as date of birth, personal identification number, etc.), making up scenarios to extract information over the phone is still a very common and applicable attack. Preparation for situations that may fall outside the edges of the scenario will further improve the success rate of the attack.

5.2.2 Convincing that the attacker is a trustworthy source

This method became famous nowadays under the name of "phishing" and usually happens over e-mail. The attacker makes the victim believe that he/she is coming from a trustworthy source or a source that its authority should not even questioned. For instance, if the attacker wants to convince a victim that a message is from the IT department of a bank, he/she can make use of the format or the template that is used by the bank's earlier messages and change the links inside the message to point to other sites, that are designed to collect victim's private information. The aim of the attacker is to force the victim to give out private information or to make an error such as clicking on a fake website, install a virus to the system.

5.2.3 Using a Trojan Horse

Software that looks like harmless but actually spies on or harms the system is called Trojan horses. The main difference of this kind of software from self-infecting viruses or worms is their dependency on the user itself to spread. Trojan horses can be installed on the system from untrusted sources, downloaded files in the disguise of a well-known application, peer-to-peer sharing networks or automatically by computer viruses.

Another way of spreading the Trojan horses is the “road apple method” which is leaving media having the Trojan on purpose horse somewhere that will be visible by the victim and will look like it was thrown away or forgotten instead of using conventional ways such as e-mail or the web. Once the media gets the victim’s attention and victim connects or inserts the media to his/her computer, Trojan horse runs and installs itself onto the system.

5.2.4 Offering help, money, gifts etc. in the exchange of certain information

An attack aimed at personal weaknesses of the victim to access private information. In this case, the victim is convinced that he or she will benefit in the end such as filling a questionnaire asking for his/her password in the exchange of a gift or some amount of money. Convincing the user to give out his/her password to fix a problem on the system also falls into this category.

5.2.5 Getting information by gaining trust

This method is based on gaining trust of the victim by the attacker inside or outside work environment and making the victim give out private information or doing a favor willingly due to this falsly formed trust. The attacker may approach the organization as a service provider and form a trust relation with a personel that has non-trivial access to the system, abuse an already formed outside-of-work relationship, or may even create the illusion that the attacker and the victim share common interests in life to gain trust.

5.2.6 Other methods

In addition to all methods described above, there are other known information collection techniques making use of typical mistakes by organizations or employees themselves.

These include but are not limited to;

1. **Shoulder surfing:** Watching over a person's shoulder while he or she types a password or accesses a restricted system,
2. **Garbage poking:** Analyzing stuff in garbage such as disks, CDs, post-its, notes etc. for possible sensitive information,
3. **Tampering with old hardware:** Examining old hardware such as junk, second hand or donated PCs, hard drives etc. for sensitive information.

Methods for social engineering attacks cannot be limited to lists. They are only limited to the attacker's persistence and creativity. And if one considers common fraud methods, number of different attacks will increase drastically.

5.3 Threats

In the case of success, social engineering attacks pose various threats. These threats can be classified as follows:

1. **Unauthorized access:** The attacker may obtain information that is necessary to access the system. Even a mistakenly given away user password can make that happen.
2. **Service theft:** With the obtained password, the attacker may download restricted files or can utilize limited resources such as bandwidth, CPU time or disk space without permission.
3. **Loss of reputation and trust:** An organization that is a victim of a social engineering attack my lose reputation against its customers and general public. Regaining this trust and reputation is usually much harder and costly than taking precautions.
4. **Distributed denial of service:** Captured system and resources may be used to capture or harm other systems and resources. These also lead to other attacks indirectly, where in this case the origin of the attack is also another victim.
5. **Loss of data and expose of sensitive information:** In case of success, the attacker may gain access to organization's and its customers' data. The attacker

than may sell this information, use it against the organization or utilize it to craft other attacks pointed at customers. If the attacker only wants to harm the organization, he or she can simply deny access to sensitive data or make it impossible to retrieve by deleting or encrypting.

6. **Being subject to legal sanction:** Not taking necessary precautions to protect non-disclosure and privacy agreements between the organization and its customers and affiliates may cause legal action to be taken against the organization.

5.4 Precautions

Precautions against social engineering attacks are similar to those against other cyber-attacks. The definitive properties of these precautions are their extended coverage going beyond not just the computer systems and network infrastructure but also environmental measures and regular trainings.

5.4.1 Physical Security

When reviewing system security, any security holes requiring local or console access are considered unlikely regardless of their impact and risk. However, while calculating this probability of likeliness other factors such as trustworthiness of the people having access to these systems should be taken into account in addition to the security of computer systems themselves. In certain situations, precautions against high physical risks should be taken. In systems with various user profiles that all have access, user security policies should be more restrictive and strict supervising should be performed with certain permission definitions for each user group/profile.

5.4.2 Effective Security Policies

The security policies that the organization adopted should be clear, concise, logical, accessible, inclusive and applicable. Policies that are not accessible or are lacking, blurry, hard to implement in real life are destined to be ignored and to die. The trust level between the organization and employees should be determined clearly

since they are the ones who will play an active role in the application of these security policies. The sweet spot is hard to find though since a low level of trust would affect the loyalty of the employee where as an unnecessarily high level of trust would put an unnecessary burden on the employees on possible attacks.

5.4.3 Training and Enforcements

Security policies are only as valuable as the knowledge of the employees about them. For this reason, continuous training and information sessions about the policies are essential for the continuity and security of the system. In addition to this, higher management should supervise the policies closely and be willing to impose sanctions in case of any violations of the policies. Trainings and sanctions together, help employees to adopt and obey security policies.

5.4.4 Incident response

It is especially important to determine what to do in case of a social engineering attack. Many security holes can be prevented by simply reviewing existing processes. Operations like determining the actually origin for e-mails, how to treat suspicious e-mails and how to verify origins for web addresses should be written down and included in regular processes for users. Since social engineering attacks usually exploit user weaknesses, the user under attack may not realize he or she is being attacked, or in some cases even if the users realizes there has been a successful attack, he or she may not report this incident for concerns about his/her trustworthiness. To prevent situations like this, all necessary grounding should be laid out to report incidents with predetermined rules before the attack occurs.

5.4.5 Supervision and Control

Social engineering, as a concept is constantly moving, being updated and changing due to its nature so all security policies adopted to prevent such attacks should be revised and updated regularly with effectiveness checks. Supervision and

control can simply be done via drills that simulate a possible attacker and measure users' responses. Social engineering attacks are modeled in four steps:

1. Information gathering
2. Establishing a relationship
3. Abuse
4. Access

At information gathering and relationship establishment steps, subjects that can initiate conversations are determined. Most commonly, passive information collection provides enough intelligence to proceed through these steps. Organizational web sites, search engines, newsgroups, forums, job seeking sites, social networks and even yellow pages provide more than enough information about the organization, its employees, and the organizational structure within the organization. Management should focus on how much sensitive information is available on these channels and whether this can be prevented.

There are many ways to gain physical access such as acting as an employee (by means of a fake id, getting inside of the start-end of the day crowd, etc.), sneaking behind employees to get past controlled doors (such as doors that open with a keycard), acting as a guest or a provider (like a postman or technician), and trying to access the facility outside of regular work hours where different security measures apply. In addition to all these, all above mentioned social engineering methods can be used. Since these kinds of control mechanisms trail around legal edges, they should be well designed and approved.

After gaining physical access, or just to access information, various methods are applicable. Inspectors usually stalk employees (eavesdropping, shoulder surfing etc.), poke trash cans in the office, peek under keyboards, calendars, agendas, post-its, notes and publicly visible boards, fuss with unlocked computers, try to convince users to allow access to their terminals or act as an IT personnel to trick users into sharing their passwords.

Some methods that can help protecting users from these kinds of social engineering attacks are below:

- Emphasis on physical security
- User education
- Two-factor authentication mechanisms such as USB or Smartcards
- Biometric authentication systems those are hard to replicate or steal
- Strong supervision of all security practices

A prototype that makes use of speaker identification via voice recognition that is a simple demonstration of biometric authentication systems is developed and explained in the following chapter.

CHAPTER SIX
DEVELOPING A VOICE VERIFICATION SYSTEM AGAINST VOIP SOCIAL ENGINEERING ATTACKS

6.1 Speech Recognition

Speech recognition is harder to perform than speech generation. Speech recognition is an easy task for a normal human brain but to do it in the digital world, with computers is a totally different story. Although computers are very efficient at storing and processing large amounts of data they need certain algorithms and data structures to obtain meaningful results from the data they process. It is a relatively easy task for a computer to send your monthly electric bill but making a computer recognize you from your voice is not for faint-hearted (Beigi, 2011).

6.2 Speech Representation

The speech signal and some of its characteristics can be represented in two different domains: time and frequency.

Short term characteristics of a speech signal, that are the characteristics in a relatively short time span, are stationary whereas long term characteristics of the signal are indeed varying to reflect the changes in the voice of the speaker.

To evaluate and process these characteristics, voice data should be in a suitable representation. This representation can be either in time domain or frequency domain.

6.3 Feature Extraction

Acquiring the acoustic properties of the speech signal is called as *feature extraction*. Feature extraction is used in both training and recognition phases. It includes following steps:

- Frame Blocking
- Windowing

- FFT (Fast Fourier Transform)
- Mel-Frequency Warping
- Cepstrum (Mel Frequency Cepstral Coefficients)

Output of this operation is forwarded to the actual speech processing system. The main goal of feature extraction is to accurately represent the captured voice data in a concise but meaningful manner to represent the core characteristics and properties of the captured voice data. Figure 6.1 shows the flow of feature extraction steps.

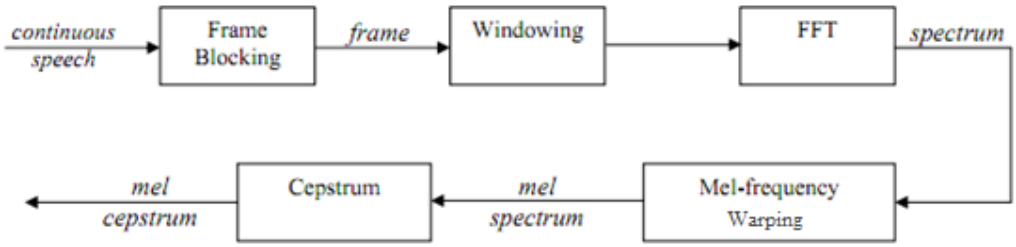

Figure 6.1 Feature extraction steps

6.3.1 Frame Blocking

Investigations show that speech signal characteristics do not change much in a sufficiently short period of time interval. For this reason, speech signals are processed in short time intervals. It is divided into frames with sizes generally between 30 and 100 milliseconds (Ursin, 2002) where each frame overlaps with the previous frame by a predefined size. This overlapping smoothens the transition from frame to frame.

6.3.2 Windowing

Next comes the windowing phase of the frames. This is to eliminate discontinuities at the end points of each frame. Let the windowing function to be

defined as $w(n)$, for $n \in [0, N)$ where N is the number of samples in each frame, the resulting signal will be; $y(n) = x(n) \cdot w(n)$. Hamming windows are preferred for most applications.

6.3.3 Fast Fourier Transform (FFT)

The next step is to take Fast Fourier Transform of each frame. This transformation is a fast way of Discrete Fourier Transform and it changes the domain from time to frequency.

6.3.4 Mel Frequency Warping

The speech signal consists of tones with different frequencies. For each tone with an actual Frequency, f, measured in Hz, a subjective pitch is measured on the 'Mel' scale. The human ear comprehends the frequencies non-linearly. Researches show that the scaling is linear up to 1 kHz and logarithmic above that (Furui, 2008). The mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following formula to compute the mels for a given frequency f in Hz (Rashidul, Jamil, Rabbani, Rahman, 2004):

$$mel(f) = 2595 * log10(1 + \frac{f}{700})$$

One approach to simulating the subjective spectrum is to use a filter bank, one filter for each desired mel-frequency component. The filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant mel-frequency interval. The Mel-Scale (Melody Scale) filter bank which characterizes the human ear perceiving of frequency is as shown in Figure 6.2. It is used as a band pass filtering for this stage of identification. The signals for each frame is passed through Mel-Scale band pass filter to mimic the human ear.

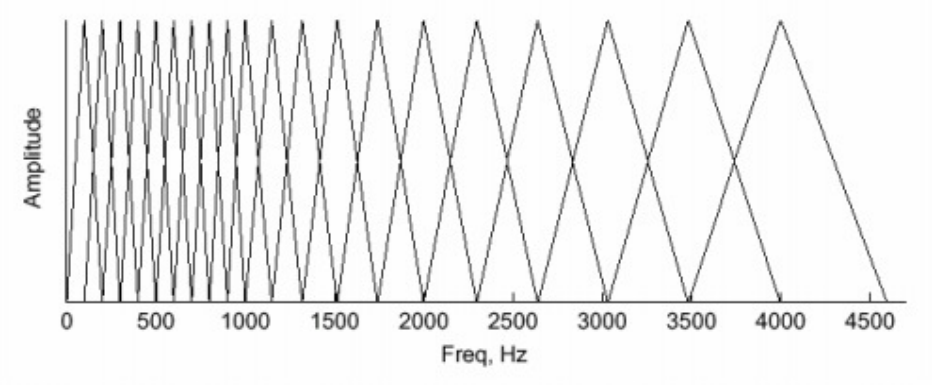

Figure 6.2 Mel scaled filter bank

6.3.5 Cepstral Coefficients

As of the final step, each frame is inverse Cosine Fourier transformed to take them back to the time domain. As a result of this process, *Mel-Frequency Cepstral Coefficients* are obtained (Hasan, Jamil, Rabbani, Rahman, 2004). These coefficients are called *feature vectors*. They are also called observation vectors in the speech recognition terminology.

6.4 Hidden Markov Model

Hidden Markov Models (HMM) are the most widely used technique in modern speech recognition systems. This is due to the fact that a great deal of effort has been devoted in research during 1980's and 1990's, making it very challenging for alternative methods to get even close to their performance with moderate investments.

Markov models were introduced by Andrei A. Markov and were initially used for a linguistic purpose, namely modeling letter sequences in Russian literature. Later on, they became a general statistical tool. Markov models are finite state automata with probabilities attached to the transitions. The following state is only dependent on the previous state.

Traditional Markov models can be considered as 'visible', as one always knows the state of the machine. For example, in the case of modeling letter strings, each state would always represent a single letter. However, in hidden Markov models the exact state sequence that the model passes through is not known, but rather a probabilistic function of it.

The underlying assumption of an HMM model in speech recognition problem is that a speech signal can be well characterized as a parametric random process, and the parameters of the stochastic process can be estimated in a precise and well defined manner. An HMM model is considered as a generator of observation sequences (i.e. feature vectors). In practice, only the observation sequence is known and the underlying state sequence is hidden. That is why this structure is called a Hidden Markov Model. HMM is very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of application.

6.5 Mathematical Understanding of Hidden Markov Model

To understand Hidden Markov Model, one first needs to understand the Markov Model. Markov Model uses the term "states" for observable events. For instance for a weather prediction model these states might be “rainy weather”, “dry weather”, “high atmospheric pressure” and “low atmospheric pressure” which are all observable states. The model makes use of “transition matrices” which indicate the transition probabilities from a certain state to another state which can be the same as the previous state in time. The ideal model would make use of all the previous states and state changes which is very compute and data intensive. A certain, simplified case of a discrete and first order approach would yield in a much simpler calculation which is only based on the previous state (Resch, 2000). If we have N distinct states, denoted by S_1 to S_N, discrete time steps denoted by t with positive integers and if we denote the actual state at a given time t as q_t we can simply say (Uchat, 2006):

$$P\left[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \dots\right] = P\left[q_t = S_j | q_{t-1} = S_i\right]$$

We can go even further and say that each state change is time independent and only depends on the previous state, which gives us

$$a_{ij} = P\left[q_t = S_i | q_{t-1} = S_j\right] \; 1 \leq i, j \leq N$$

For $a_{ij} \geq 0$ and $\sum_{j=1}^{N} a_{ij} = 1$ which is called an observable Markov Model since each state is an observable state and the system as a whole is composed of all these states and the probability of state changes.

6.6 Extension to Hidden Markov Model

In above model each state corresponds to an observable event. This model is too restrictive to be applicable to many problems of interest. We now extend this idea where observation is a probabilistic function of state. To get clear understanding we take the weather example with a different approach (Venkataraman, 2001).

Weather Model: You are locked in a room for several days, and asked about the weather outside. The only piece of evidence you have is whether the person who comes into the room bringing your daily meal is carrying an umbrella or not.

The probability that your caretaker carries an umbrella is 0.1 if the weather is sunny, 0.8 if it is actually raining, and 0.3 if it is foggy as listed in Table 6.1.

Table 6.1 Probability of carrying an umbrella based on the weather

Weather	Probability of Umbrella
Sunny	0.1
Rainy	0.8
Cloudy	0.3

The equation for the weather Markov process before you were locked in the room was:

$$P(q_1, \dots, q_n) = \prod_{i=1}^{n} P(q_i|q_{i-1})$$

However, the actual weather is hidden from you. Finding the probability of a certain weather $q_i \in \{\text{sunny, rainy, cloudy}\}$ can only be based on the observation x_i, with $x_i = umbrella$, if your caretaker brought an umbrella on day i, and $x_i = no\ umbrella$ if the caretaker did not bring an umbrella. This conditional probability $P(q_i|x_i)$ can be rewritten according to Bayes' rule:

$$P(q_i|x_i) = \frac{P(x_i|q_i)P(q_i)}{P(x_i)}$$

or, for n days, and weather sequence $Q = \{q_1, \dots, q_n\}$ as well as 'umbrella sequence'

$$X = \{x_1, \dots, x_n\}$$

$$P(q_1, \dots, q_n|x_1, \dots, x_n) = \frac{P(x_1, \dots, x_n|q_1, \dots, q_n)P(q_1, \dots, q_n)}{P(x_1, \dots, x_n)}$$

Using the probability $P(q_1, \dots, q_n)$ of a Markov weather sequence from above, and the probability $P(x_1, \dots, x_n)$ of seeing a particular sequence of umbrella events (e.g., {umbrella, no umbrella, umbrella}).

The probability $P(x_1, \dots, x_n|q_1, \dots, q_n)P(q_1, \dots, q_n)$ can be estimated as $\prod_{i=1}^{n} P(x_i|q_i)$ if you assume that, for all i, the q_i, x_i are independent of all x_j and q_j, for all $j \neq i$.

We want to draw conclusions from our observations (if the person carries an umbrella or not) about the weather outside. We can therefore omit the probability of seeing an umbrella $P(x_1, \dots, x_n)$ as it is independent of the weather, that we like to

predict. We get a measure for the probability, which is proportional to the probability, and which we will refer as the likelihood L.

$$P(x_1, \dots, x_n | q_1, \dots, q_n)$$

$$L(q_1, \dots, q_n | x_1, \dots, x_n) = P(x_1, \dots, x_n | q_1, \dots, q_n) \cdot P(q_1, \dots, q_n)$$

With our (first order) Markov assumption it turns to:

$$P(x_1, \dots, x_n | q_1, \dots, q_n)$$

$$L(q_1, \dots, q_n | x_1, \dots, x_n) = \prod_{i=1}^{n} P(x_i | q_i) \cdot \prod_{i=1}^{n} P(q_i | q_{i-1})$$

Suppose the day you were locked in was a sunny day. The following day, your caretaker came with an umbrella into the room. You would like to know, what the weather was like on this second day. First we calculate the likelihood for the second day to be sunny. Respectively, sunny, rainy and cloudy represented by s, r, c in the equations below.

$$L(q_2=s|q_1=s, x_2=\text{umbrella}) = P(x_2=\text{umbrella}|q_2=s) \cdot P(q_2=s|q_1=s)$$

$$=0.1 \cdot 0.8 = 0.08$$

Then for the second day to be rainy:

$$L(q_2 = r|q_1 = s, x_2 = umbrella) = P(x_2 = umbrella|q_2 = r) \cdot P(q_2 = r|q_1 = s)$$

$$=0.8 \cdot 0.05 = 0.04$$

And finally for the second day to be cloudy:

$$L(q_2 = c|q_1 = s, x_2 = umbrella) = P(x_2 = umbrella|q_2 = c) \cdot P(q_2 = c|q_1 = s)$$

$$=0{,}3 \cdot 0.15 = 0.045$$

Thus, although the caretaker did carry an umbrella, it is most likely that on the second day the weather was sunny.

6.7 Implementation of the System

In this thesis MATLAB software is used to record and analyze speech data. Also two MATLAB toolboxes, provided under GUI are used.

- VOICEBOX is a toolbox which consists of many functions for speech processing. Especially for Mel-Frequency Cepstral Coefficients (MFCC) calculation, MELCEPST function is used.
- H2M is a different toolbox than VOICEBOX which includes a number of functions implementing the Hidden Markov Model.

Two MATLAB functions which are “testing main, training main" are developed. Total ten numbers, which are “Zero, one, two, three… nine”, are used for this speech recognition system.

6.8 Voice Recording

Firstly, user has to complete a voice recording session. After the recognition, recorded words will be operated in training stage to get HMM.

Every word is recorded as a “.wav” file with 22050 Hz sampling frequency.

6.8.1 Training

During the training process, roughly voice recording of a word is transformed into HMM. The processing is needed to create a folder which includes HMM of words. In testing process all of the HMM will be used to obtain the highest score to recognize a word.

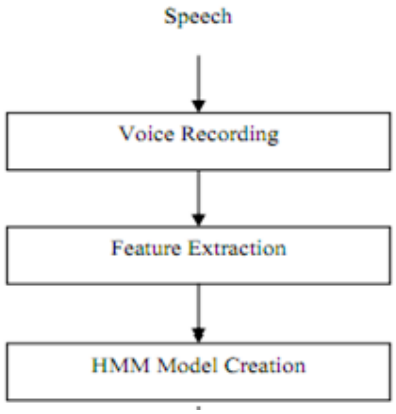

Figure 6.3 General block diagram of training stage

In the design of most speech recognition systems a key assumption is that the segment of a speech signal can be considered as stationary over an interval of few milliseconds. Therefore the speech signal can be divided into blocks which are usually called frames. The spacing between the beginnings of two consecutive frames is in the order of 10 msecs, and the size of a frame is about 25 msecs (Furui, 2008). That is, the frames are overlapping to provide longer analysis windows. Within each of these frames, some feature parameters characterizing the speech signal are extracted.

Before creating the HMM models of the words recorded, a feature extraction process is applied to get the observation vector from the sound data. In this thesis, FFT based MFCC is chosen as the feature extraction method since it was already provided by the Voicebox toolbox and was proved to be useful by previous studies (Jang, 2012)

After the feature extraction phase, the model parameters are initialized by using hmm_mint function provided by H2M toolbox which simply “chops” each parameter into blocks where the number of blocks is equal to the states of the HMM. Following this, “Left-Right HMM” path is followed as explained in H2M’s documentation (Cappe, 2001). hmm_mest and mix_par functions are used to extract mean vectors(mu) and covariance matrices(sigma) which are then stored on the disk for the

testing phase. Note that hmm_mest function assumes that the Markov chain starts from the initial state which is a valid assumption for voice recognition purposes.

6.8.2 Testing

In this process an unknown word is to be recognized by using created HMM in training stage. General block diagram of testing stage is given Figure 6.4.

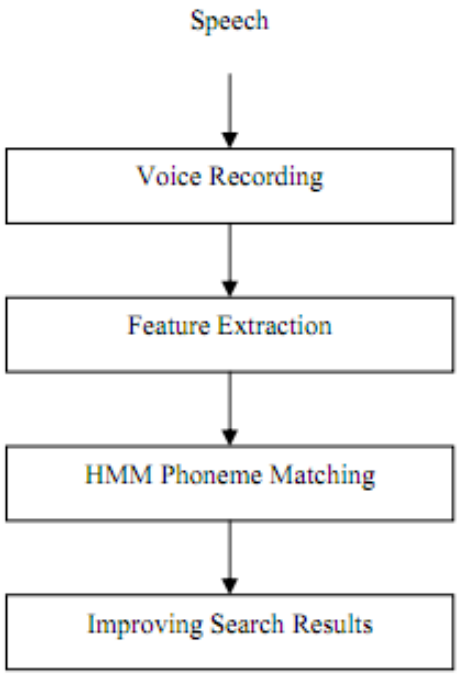

Figure 6.4 General block diagram of testing stage

The system first generates a random number between 0 and 9 and displays it to the user. It also uses this number to select the subset of the voice data it will use to match against, along with the phone number provided.

The first two steps performed in recognition are the same as the training phase: voice is recorded, feature extraction and HMM model generation is performed. Following these, a “score” is calculated for each voice data available by the given constraints. This scoring phase makes use of the well-known Viterbi DP algorithm by calling hmm_vit function in H2M toolbox. The likelihood score is an exponent

and it is calculated using the stored mu and sigma values for each sample against the extracted observation vector from the new recording.

This operation creates an array of scores that correspond to each sample in the which represent different users. The maximum score is the most likely user that is recognized but since this is a voice authentication system, some further filtering is necessary. For instance, the score should be above then a certain empirical threshold to prevent low scores from identifying a user. Also, the top two scores should at least have a significant difference between them for accurate user identification. If the scores are too close to each other, it means the system could not differentiate these two users from each other accurately. This also causes continuous false negatives if the same user has more than one sample under different names since the system thinks it is confusing two different users with each other.

6.9 Sample Training and Verification Session with Screenshot

6.9.1 Main Menu

In this part, a sample training and recognition session is given from the speech recognition application with screenshots. The application is developed in Turkish. In Figure 6.5 main menu of the application is seen. Then, other parts of the application are selected through this menu.

For training a user must enter his/her name and a phone number. The speech recognition system will be keep user recorded voice with phone number. After entering of the name and phone number, “Next” button should be used for starting training.

6.9.2 Training

The Graphical User Interface (GUI) of the training part of the application is shown below. In the training screen, the user is instructed to record his/her voice by speaking numbers appearing at the top of the screen.

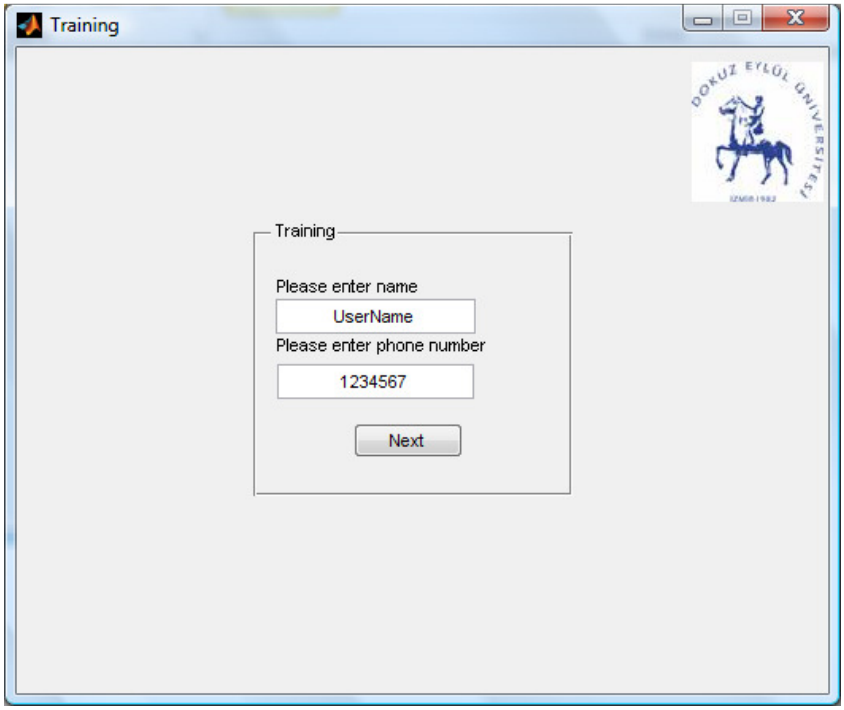

Figure 6.5 Main menu of the speech recognition system

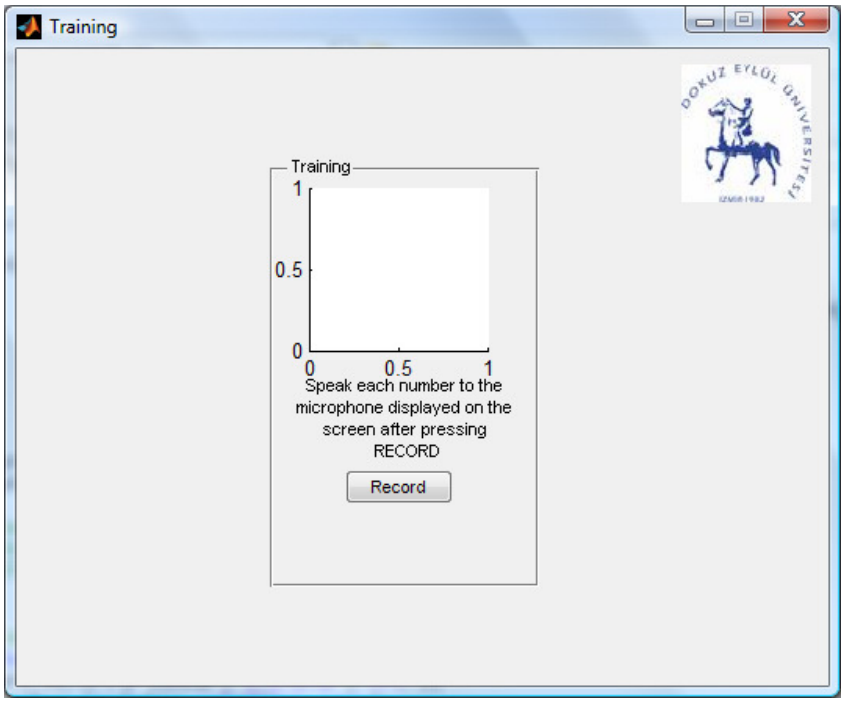

Figure 6.6 Training menu of the speech recognition application

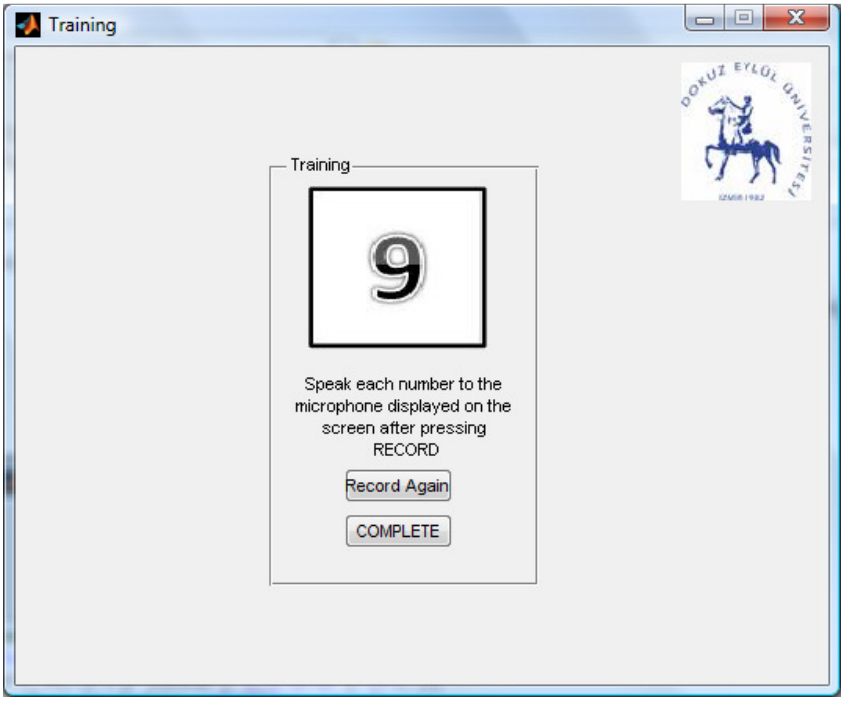

Figure 6.7 Training menu of the speech recognition application

After all numbers from zero to nine are recorded, the training is completed by pressing “COMPLETE” button with the creation of words’ HMM models. After this procedure, the system is ready for recognition.

If the user want to record her/his voice again or a new user will be introduced to recognition system, “Record Again” button can be used.

6.9.3 Verification

The verification part of the application is shown below. Firstly, user should enter his/her phone number. This step will be eliminated once the system is integrated into the core VoIP service provider infrastructure.

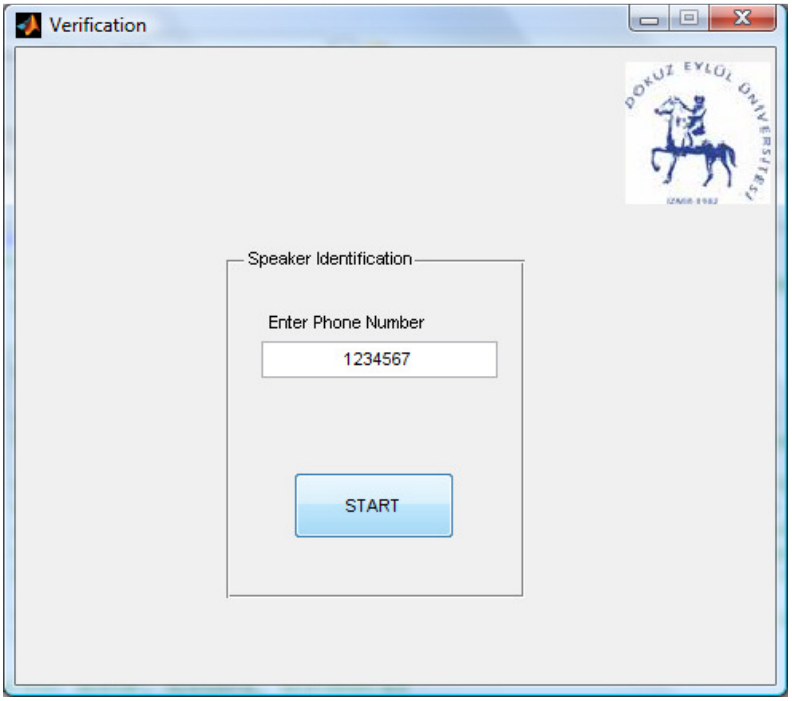

Figure 6.8 Verification menu of the speech recognition application

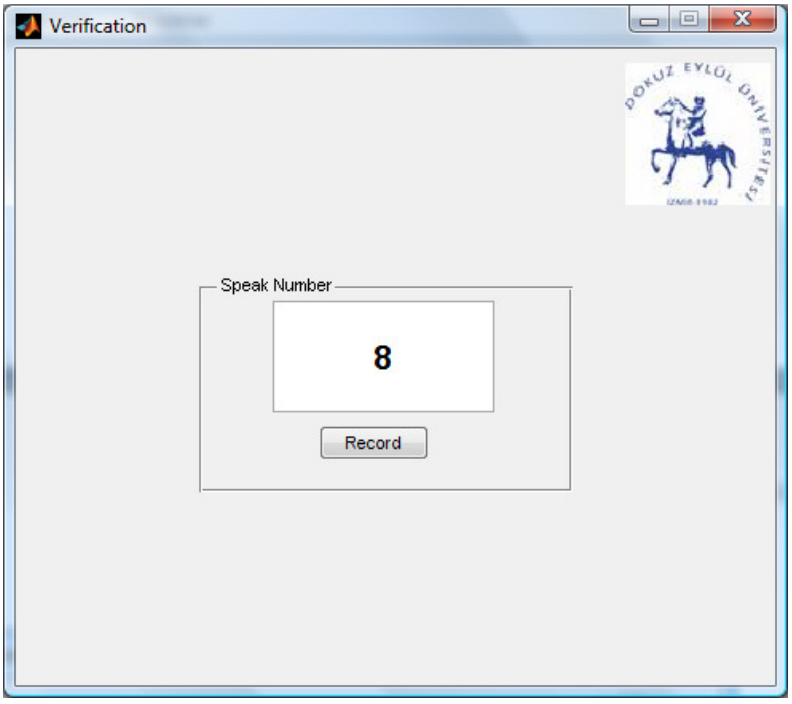

Figure 6.9 Displayed random number for verification system

A random number between zero and nine will be displayed to the user for him/her to pronounce. The system then, compares the properties of this newly recorded sound with the ones previously recorded for this same phone number. In a production quality system, ten numbers will obviously not be enough. An ideal system would generate sentences or at least two or more random words for better accuracy and confidence.

The system extracts the properties of the newly recorded voice data as it does in the training section and then calculates a score against all the previously recorded voice data for that same telephone number and the displayed number. The highest scoring user is selected to be the identified user though this highest score should be above a certain threshold value (-36000) and should have a difference (200) greater or equal with its closest competitor. This ensures better accuracy and prevents false positives though these threshold and difference limits should be calculated for each different word rather than being constant for all words and numbers in a production level system.

Once the user is identified by satisfying all the conditions explained above, the name for the user is displayed in the box with a green background to indicate the user is identified properly.

If the user is not identified then “Cannot verify speaker” message in the box with a red background will be shown.

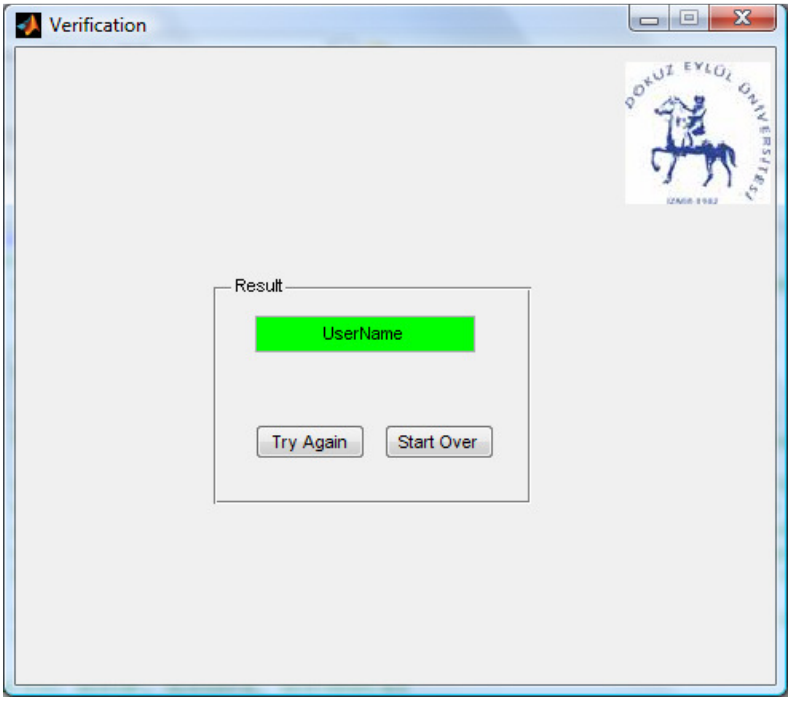

Figure 6.10 Successful verification result screen

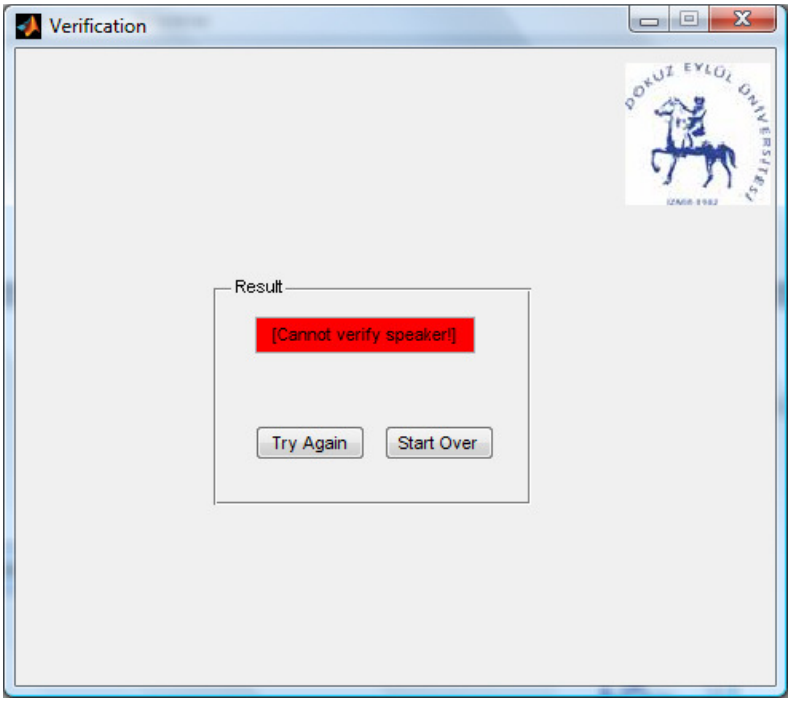

Figure 6.11 Unsuccessful verification result screen

6.9.4 Results

6.9.4.1 Detailed, Gender Specific Analysis

For testing of the system, all 10 numbers were trained and a total of 200 HMM models were created for 20 users. Ten female and ten male speakers are used to generate the reference templates. Each trained number's sample is taken from speakers for performance testing.

Voice verification prototype was performed 20 times with 10 males and 10 females. Results are as below;

Table 6.2 Voice authentication prototype performance result with 10 males and 10 females

	Number of Trials	True Positive rate %	True Negative rate %	False Negative rate %	False Positive rate %
10 Male	20	92	90	8	10
10 Female	20	96	94	4	6

To be more precise, in Table 6.2, true positive performance represents the rate of the user who is registered in the system and identified correctly; true negative column shows the user who is not registered in the system and cannot be identified by the recognition system, as expected. These two ratios mean that the system is working correctly. The third column, false negative, indicates the rate of failed attempts for a user who is registered on the system but not recognized. Last column, false positive, represents the ratio of the users who do not exist in the system but was identified as a valid user by the system. It means system confused the non-existent user with one of the previously registered user.

True positive and true negative rates may also be considered as correct results ratio which indicates the total success ratio of the voice authentication prototype is pretty high as indicated above in Table 6-2. This high accuracy does not change the fact that false positives are unacceptable for such a system and should be corrected.

The reason of the relatively lower accuracy for the male users is assumed to be the similarity between the dictation styles but no additional study to back this assumption is made since it is not the main focus of the study.

6.9.4.2 Recognition Ratios Separated by Numbers

In a different, more use case oriented set of tests without any gender distinction, with the focus being only on the correct recognition rates of the spoken numbers results are calculated and presented in Table 6-3. In this case, same number was shown 50 times and speaker was chosen randomly.

Table 6.3 Voice authentication prototype success rates per each number used in recognition system

Test Number	Number of Trial	Number of successful recognition	Success Rate %
0	50	45	90
1	50	46	92
2	50	48	96
3	50	41	82
4	50	46	92
5	50	43	86
6	50	47	94
7	50	46	92
8	50	46	92
9	50	47	94

As seen in Table 6.3, success recognition rates are very close to each other except for the numbers 3 and 5. The reason of the low recognition rate for the number 3 and number 5 is that these numbers contain Turkish characters. Studies with Turkish characters show low success rate in speech recognition system. (Salor, Pellom, Çiloğlu, Hacıoğlu, Demirekler, 2002).

Also Creutz, M., Hirsimaki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arısoy, E., Saraçlar, M., Stolcke, A., (2006) define the reason of low recognition rate for morphologically rich language like Turkish as below;

As automatic speech recognition systems are being developed for an increasing number of languages, there is growing interest in language modeling approaches that are suitable for so-called "morphologically rich" languages. In these languages, the number of possible word forms is very large because of many productive morphological processes; words are formed through extensive use of inflection, derivation and compounding (such as the English words 'rooms', 'roomy', 'bedroom', which all stem from the noun 'room'). English is a fairly "morphologically poor" language, and consequently language modeling on the word level has proven successful, or at least satisfactory.

The recognition vocabulary consists of a list of word forms observed in the training text, and n-gram language models are estimated over sequences of words. In comparison to other "morphologically richer" languages, for English this approach does not typically lead to huge vocabularies or severe scarcity problems, nor to a high proportion of out-of-vocabulary (OOV) words in a held-out independent test text. A high OOV rate would be problematic, since OOVs are words that are not present in the recognition vocabulary, and thus can never be recognized correctly by the speech recognizer. In addition, the presence of out-of-vocabulary words may cause the misrecognition of several in-vocabulary words adjacent to the OOVs. Common wisdom, which is supported by experiments, says that each OOV word causes between 1.5 and 2 word errors, on average.

CHAPTER SEVEN
CONCLUSION

In this thesis a comprehensive comparison between traditional PSTN and VoIP networks is performed. All underlying protocols of VoIP are examined thoroughly and their technical weaknesses in terms of security are tried to be covered with the help of many other studies that already exists in this area.

Many technical vulnerabilities can be mitigated by means of already established strong encryption mechanisms that are available to modern computer networks and protocols. Also other means such as firewalls, network policies and similar can be used to prevent almost all outstanding issues on this area.

In addition to the existing technical attack analyses, social engineering attacks are also analyzed with the help of relatively small number of existing studies, especially in the VoIP field. After these analyses, it is determined that biometrics, which is a trending and commonly used method for authentication and identity verification is found suitable also for VoIP networks and a voice based biometric speaker identification system is prototyped for demonstrative purposes using Hidden Markov Models method, which is a very common method used in this field.

The prototype proved useful and identified users at relatively high accuracy. But it is also observed that there are still false negatives, which are frustrating for the end-users and false negative that should never exist in a production-ready system. According to many research out there on the biometrics field, it has been proven that 100% accuracy, especially for false positives is not an achievable goal using a single biometric unit (voice in this case) at a reasonable cost and efficiency. Thus it is decided that using multiple biometric vectors together to identify a person (Beigi, 2011) would be the best possible solution.

In addition to randomized word reading based voice authentication, an additional finger-print verification, face recognition, palm vein pattern matching, retinal scan or

similar and even a traditional “password” would increase the systems accuracy and reliability considerably. Further work on the subject should be conducted on a more generalized speaker-dependent voice recognition including the ability to recognize many words with less training, learning from each successful verification, and deployment of this kind of verification systems to large production environments since this paper only cracks the door open for biometric authentication mechanisms on VoIP networks.

REFERENCES

Allan A., (2004). *Security and Privacy.* Security Administration

Anderson R.J., (2000). *Security Engineering*. ISBN 0-471-38922-6

Anwar, Z., Yurcik, W., Johnson, R. E., Hafiz, M., and Campbell, R. H. (2006). Multiple Design Patterns for Voice over IP (VoIP) Security. In Proceedings of the *IEEE Workshop on Information Assurance (WIA)*, held in conjunction with the 25thIEEE International Performance Computing and Communications Conference, (IPCCC)

Barbieri, R., Bruschi, D. (2005). *Voice over IPSec: Analysis and Solutions.* Milano: Dipartimento di Scienze Dell'Informazione Universita degli Studi

Beigi, H. (2011). *Speaker Recognition, Biometrics.* Jucheng Yang (Ed.), ISBN: 978-953-307-618-8, InTech

Bhan, S., Clark, J., Cumeo, J., Ramirez-Mejia, J. (2006). *Information Security Issues in Voice Over Internet Protocol.* Georgia Tech College of Engineering ECE

Bonneau J., (2012). *The science of guessing analyzing an anonymized corpus of 70 million password.* Master Thesis. Computer Laboratory University of Cambridge

Burr W.E., Dodson D.F., Polk W.T. (2006*). Information Security*. National Institute of Standards and Technology

Butcher, D., Li X., Guo J. (2007). Security Challenge and Defense in VoIP Infrastructures. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 37*(6):*1152–1162*

Cappe, O., (2001). *A Set of Matlab / Octave Functions for the EM Estimation of Mixtures and Hidden Markov Models*, ENST dpt. TSI / LTCI (CNRS- URA 820), Paris, France

Çakır, C., Kaptan, H. (2009). *VoIP Teknolojilerinde Opnet Tabanlı Güvenlik Uygulaması*. Master Thesis, Department of Computer Engineering, Marmara University

Cao, F. and Malik, S. (2006). Vulnerability Analysis and Best Practices for Adopting IP Telephony in Critical Infrastructure Sectors. *IEEE Communications Magazine, 44*(4*):138–145*

Creutz, M., Hirsimaki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arısoy, E., Saraçlar, M., Stolcke, A., (2006). *Analysis of Morph-Based Speech Recognition and the Modeling of Out-of-Vocabulary Words Across Languages*. Master Thesis, Helsinki University of Technology, SRI International / International Computer Science Institute

Chen, T.M., (2009). *From Circuit Switched to IP-based Networks.* Texas: Department of electrical Engineering

Cooke, M., (2006). *Hidden Markov Models and Beyond*. Master Thesis, University of Sheffield

Drew, P., Gallon, C. (2003). *Next Generation VoIP Network Architecture* Retrieved November, 2012 from http://www.msforum.org

Errol, A. B. (2007). *Network Security: VoIP Security on Data Network.* USA: Information Security Curriculum Development Conference'07

Florencio, D., Herley, C., (2007) *A Large- Scale Study of Web Password Habits.* Bonff , Canada

Furui, S., (2008). *Speaker Recognition.* Scholarpedia, 3(4):3715, Tokyo Institute of Technology

Gales, M., Young, S., (2007). *The Application of Hidden Markov Models in Speech Recognition*, Foundations and Trends in Signal Processing Vol.1 No:3

Garg, S., Singh, N., Tsai, T. (2005). *Schemes for Enhancing the Denial of Service Tolerance of SRTP*

Hasan, R., Jamil, M., Rabbani, G., Rahman, S., (2004). *Speaker Identification Using Mel Frequency Cepstral Coefficients*, ICECE, Dhak Bangladesh ISBN 984-32-180-4

Homayoon, B., (2011). *Speaker Recognition*. USA

Jang, R., (2012). *Audio Signal Processing*. Retrieved November, 2012 from http://neural.cs.nthu.edu.tw/jang/books/audiosignalprocessing/speechfeaturemfcc.asp?title=12-2%20mfcc

Juang B.H., Rabiner L.R., (2007). Hidden Markov Models for Speech Recognition, *Technometrics, Vol.33, No:3 pp. 251-272*

Lawecki, P., (2007). *VoIP Security in Public Networks.* Master Thesis, Chair of Telecommunication and Computer Networks, Poznan University of Technology & Institute of Communication Networks and Computer Engineering, University of Stuttgart

Masuko, T., (2002). *HMM-Based Speech Synthesis and Its Application.* Retrieved November, 2012 from http://citeseerx.ist.psu.edu

Meggelen, J. V., Smith, J., Madsen. L., Asterisk (2007). *The Future of Telephony.* Beijing Cambridge

Orrblad, J. (2004). *Alternatives to MIKEY / SRTP to secure VoIP.* Stockholm: KTH Microelectronics and Information Technology

Oosterloo, B., (2008). *Managing Social Engineering*. Master Thesis, University of Twente

Patel, I., Rao, Y.S., (2010). Speech Recognition Using HMM with MFCC-An analysis using frequency spectral decomposition technique. *Signal & Image Processing : An International Journal (SIPIJ) Vol:1, No:2*

Patrick, C.K. Hung., Miguel Vargas Martin (2006). *Security Issues VoIP Applications.* Ottawa: IEEE CCECE/CCGEI

Paul, D.B., (1990). *Speech Recognition using Hidden Markov Models*. The Lincoln Laboratory

Paul, E. Jones (2004). *Presentation describing features and characteristics of H.323.* Rapporteur, ITU-T

Persky, D. (2007). *VoIP Security Vulnerabilities*. SANS Institute

Rabiner, R., (1989). A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceeding of the IEEE, Vol.77, No:2*

Resch, B., (2000). *Hidden Markov Models*, Signal Processing and Speech Communications Laboratory, Inffeldgasse 16c/11

Reynolds, B., Ghosal, D. (2003). *Secure IP Telephony using Multi-layered Protection.* In Proceedings of the ISOC Symposium on Network and Distributed Systems Security (NDSS)

Seedorf, J. (2006). Security challenges for peer-to-peer SIP. *IEEE Network, 20(5):38–45*

Solar, Ö., Pellom, B., Ciloglu, T., Hacioglu, K., Demirekler, M., (2002). *On Developing New Text and Audio Corpora and Speech Recognition Tools for the Turkish Language*, Inter. Conf. on Spoken Language Processing, Vol.1, pp. 349-352, Denver, USA

Thermos, P., Takanen, A. (2008). *Securing VoIP Networks*. Pearson Education

Uchat, N., (2006). *Hidden Markov Model and Speech Recognition*. Department of Computer Science and Engineering Indian Institute of Technology, Bombay Mumbai

Ursin, M., (2002). *Triphone Clustering in Continuous Speech Recognition*. Master Thesis, Helsinki University of Technology

Venkataraman, S., (2001). *Applications of Hidden Markov Models to Computational Biology problems*

APPENDIX A

```
function training_main(isim,sayi,ozellik)
steps = 180;
for step = 1:steps
    % computations take place here
    waitbar(step / steps)
end
    warning off all
    Fs = 22050;
    X = wavrecord(0.8*Fs,Fs); % 16 bit mono 22050 Hz
    [b,a] = butter(9,3000/11025,'low');
    y = filter(b,a,X);
    dosya=strcat(sayi,'\',ozellik,isim);
    wav_file_name = ['C:\Users\Seylan\Desktop\SEYLAN_TEZ\training_words\'
dosya '.wav'];
    wavwrite(y,Fs,16,wav_file_name);
    warning('off','MATLAB:dispatcher:InexactCaseMatch');
    numStates = 5;
    train_word(dosya,numStates);
end
function y=pattern(y)
xabs=abs(y);
[h,xx]=hist(xabs,100);
histout=y.*(xabs>=xx(14));
a=0;b=0;
for i=1:length(histout)
    if (histout(i)~=0) && a==0
        a=i;
    end
    if (histout(length(histout)-i)~=0)&&(b==0)
        b=length(histout)-i;
    end

    if (a~=0)&&(b~=0)
        break;
    end
end
y=y(a:b);
%y1=padarray(y,a,0,'pre');
%    % hist(abs(x))
%    % hold on
%        plot(k,'b');
%      hold on
%     plot(y1,'r');
%      hold on


end
function train_word( wordName, numStates )
disp(['Training word: ' wordName]);
wav_file_name = ['C:\Users\Seylan\Desktop\SEYLAN_TEZ\training_words\'
wordName '.wav']; % Every word is recorded in 22050 Hz., 16 bit, mono Wave
```

```
[y, fs] = wavread(wav_file_name);
observationVector = melcepst(y, fs, '', 22, floor(3*log(fs)),128,32);
%frame length 128 = 5.8 msec 1 sn de 22050 örnek
%frame overlap 32
%melcepst ile Öznitelik Vektörleri çıkarıldı dosyanın,, 22 boyutlu bir
%matris,satır sayısı frame sayısına eşit
%feature extraction method
%HMM Parametrelerinin Çıkarılması%%%%%%%%%%%%%%
%Mel cepst ile  fenomlar için bir cepstrum cıkarıldı%%%%%
%HHM fonksiyonları ile 2 parametre cıkardık mu ve sigma%%%
%2 row 2 states hmm and 22 column obserwatıon wector
N = numStates;
DIAG_COV = 1;
QUIET = 1;
st = 1;
X = observationVector;
T = size(X,1);

[my_mu,Sigma] = hmm_mint(X, st, N, DIAG_COV,QUIET);
Sigma = ones(N,1)*mean(Sigma);

NIT = 10;
logl = zeros(1, NIT);

A = sparse(0.85*diag(ones(1,N))+0.15*diag(ones(1,N-1),1));
A(N,N) = 1;
p = size(observationVector, 2);

for n = 1:NIT
 [tmp, logl(n), gamma] = hmm_mest(X, st, A, my_mu, Sigma, QUIET);
 [my_mu, Sigma] = mix_par(X, gamma, DIAG_COV, QUIET);
Sigma = ones(N,1)*(sum((sum(gamma)'*ones(1,p)).*Sigma)/T);
end

my_sigma = Sigma(1,:);
eval(['save C:\Users\Seylan\Desktop\SEYLAN_TEZ\hmm_model_files\' wordName
' my_mu my_sigma']);
end
```

```
function result=testing_main(ozellik,k)
    warning off all
    Fs = 22050;
    X = wavrecord(0.8*Fs,Fs); % 16 bit mono 22050 Hz
    [b,a] = butter(9,3000/11025,'low');
    y = filter(b,a,X);

wavwrite(y,Fs,16,'C:\Users\Seylan\Desktop\SEYLAN_TEZ\testing_words\test_wo
rd.wav');
    result=test_word('test_word',5,ozellik,k);
end

function res=test_word( wordName, numStates, ozellik, sayi)
wav_file_name = ['C:\Users\Seylan\Desktop\SEYLAN_TEZ\testing_words\'
wordName '.wav'];
[y, fs] = wavread(wav_file_name);
observationVector = melcepst(y, fs, '', 22, floor(3*log(fs)),128,32);
N = numStates;
A = sparse(0.85*diag(ones(1,N))+0.15*diag(ones(1,N-1),1));
A(N,N) = 1;


direction=strcat('C:\Users\Seylan\Desktop\SEYLAN_TEZ\training_words\',int2
str(sayi),'\',ozellik,'*.wav');
 B=dir(direction);
 l=length(B);
 score=zeros(1, l);
 for i=1:1:l
    fiName =strtok(B(i).name, '.');
    fileName=strcat(int2str(sayi),'\',fiName);

    modelFileName = ['C:\Users\Seylan\Desktop\SEYLAN_TEZ\hmm_model_files\'
fileName '.mat'];

    load(modelFileName);
    Sigma = ones(N,1) * my_sigma;

    %viterbi ile en yakın skor elde ediliyor
    score(i) = hmm_vit(observationVector, A, [1 zeros(1,N-1)], my_mu,
Sigma, 1);
 end
 disp(score);
 [max_score, max_ind] = max(score);

 if size(score) > 1
     score(max_ind) = min(score);  % try to eliminate self
     min_diff = max_score - max(score);
 else
     min_diff = 200;
 end

 if max_score >= -36000 && min_diff >= 200  % hard treshold
    res=B(max_ind).name;
    res=res(length(ozellik)+1:length(res)-4);  % get rid of the .wav part
```

```
else
    res = '';
end

end
```

APPENDIX B

Term	Definition
ATM	Asynchronous Transfer Mode
CLID	Call Line Identification
DCCP	Data Congestion Control Protocol
DoS	Denial of Service
FFT	Fast Fourier Transform
GSM	Global System for Mobile Communications
GUI	Graphical User Interface
HMM	Hidden Markow Model
HTTP	Hypertext Transfer Protocol
IETF	Internet Engineering Task Force
IP	Internet Protocol
IPSec	Internet Protocol Security
ISDN	Integrated Services Digital Network
ITU-T	International Telecommunications Union Telecommunication Standardization Sector
LAN	Local Area Network
MC	Multipoint Controller
MCU	Multipoint Controller Unit
MP	Multipoint Processor
NIST	National Institute of Standards and Technology
OOV	Out of Vocabulary
PCM	Pulse Code Modulation
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RFC	Request for Comment
RSV	Resource Reservation Protocol
RTP	Real-Time Transfer Protocol

RTCP	Real-Time Control Protocol
SCTP	Stream Control Transmission Protocol
SDP	Session Definition Protocol
SIP	Session Initiation Protocol
SPIT	Spam over Internet Telephony
SRTP	Secure Real-Time Transport Protocol
TCP	Transmission Control Protocol
TLS	Transport Layer Security
UAC	User Agent Client
UAS	User Agent Server
UDP	User Datagram Protocol
VLAN	Virtual Local Area Network
VoIP	Voice over Internet Protocol