APPLYING SPEAKER RECOGNITION ON VOIP AUDITING

XUAN WANG, JIANCHENG LIN

Intelligence Computing Research Center, Department of Computer Science, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China
E-MAIL: wangxuan@insun.hit.edu.cn, linjiancheng1983@163.com

Abstract:
This paper presents an approach to combining VoIP auditing and speaker recognition, and an application of using speaker recognition on VoIP voice data is described. In this paper, it deals with how to obtain VoIP voice data when audio communication: signaling protocol analysis and voice data traffic analysis. After converting the raw voice data to wave files, the speaker recognition can use them to identify the speaker and the result shows it is an effective way. This paper also discusses the influence by codec compression and packet loss of VoIP communication.

Keywords:
Speaker recognition; VoIP; VoIP auditing; SIP; RTP; Codec

1. Introduction

VoIP (Voice over Internet Protocol) is a solid technology available since some years that help people to communicate via voice using the IP protocol instead of traditional telephone lines, like other new technologies that deliver voice communications over Internet and Internet protocol networks have the potential to support innovation, improve access to communications services, and reduce the cost of ownership within many developing countries and around the world. In the past several years, VoIP was only a technology prospect for the “future works” segment of telephony and networking papers. But now, telecommunications companies and other organizations have already, or are in the process of, exploiting and building up their data networks, or even moving their services to VoIP. Total equipment purchases of VoIP gateways, soft switches such as IP Private Branch Exchange (IP PBX), and VoIP application servers are expected to reach almost $12 billion by 2006, a six-fold increase over 2001.

VoIP can be applied in Telephone Bank, or used for collaboration by two people who are between long distance. However, it can become facility for criminal. Let us look at these two scenarios:

Scenario 1: An impostor pretends to be the telephone bank’s account holder, he has the password. Then how to forbid his accessing?

Figure 1. Imposter access telephone bank by VoIP

Scenario 2: Police charges two men with terrorists, and they have obtained the terrorists’ audio chatting record. Then how to identify the speaker?

Figure 2. Terrorists plan an attack by VoIP

One way is capturing and making use of their voice. Speaker, or voiceprint, recognition is a kind of biometric technology that uses an individual’s voice for recognition purposes. Speaker recognition can be applied in intelligent user interface, forensics and surveillance. “Most of the commercialization has focused on using speaker recognition as a biometric to control access to information, services, or computer accounts. As with other biometrics, speaker recognition offers the ability to replace or augment PINs and passwords with something that cannot be forgotten lost or stolen” [1].

Before Speaker recognition, we should capture the voice data legally, that is VoIP auditing. In this paper, we only constrain VoIP auditing to legal interception when capturing voice. Legal interception is an effective means adopted by many countries to insure national security and society stabilization. In Europe, France, Germany, Italy,
Portugal and Spain permit flexible legal interception, and in Northern Ireland and England forcible legal interception are allowed.

Combining legal interception on VoIP auditing and Speaker Recognition can assist security department to investigate crime on VoIP. This application also can be “home-parole monitoring and prison call monitoring, and there has also been discussion of using this kind automatic systems to corroborate aural/spectral inspections of voice samples for forensic analysis”[1]. Cisco has developed VoIP Monitor Server 4.2, and National Chiao Tung University in Taiwan has done researched on SIP (a kind of VoIP signaling protocol) analyzer. IBM T.J. Watson Research Center[2] has presented the research on speaker recognition from compressed VoIP packet stream.

2. Speaker Recognition

Speaker recognition encompasses all the activities involving the identification of a speaker, based on his or her voice and the clustering of speakers based on similarities of their individual information included in speech waves. Potential applications of speaker recognition systems include access control to distant databases, reservation or information services, as well as banking over the telephone network. The general techniques used for the three main components of these systems, namely, front-end processing, speaker modeling, and pattern matching, are briefly described next.

2.1. Front-end Processing/Feature Extraction

Individual information is included in speech waves. The task of front-end processing is to extract the information, called feature extraction. This step involves the conditioning of the signals for further analyses. First, the speech samples would be filtered to remove noisy and the silent portions. Second, now the signal would be divided into frames which is called Windowing (A speech interval typically spans 10-30ms per frame). The extracted feature presents the original characters, even it can be used to characterize the voice. There are two categories of feature extraction: the temporal coding methods and the frequentional coding methods. The former includes LPC, LPCC, PLP, LAR, etc, while the later including FFT filter banks, Cepstr, MFCC, etc. Since the size of the feature vector is large, Principal Component Analyses is used to determine the significant dimensions of the given data.

2.2. Speaker Modeling

In speaker modeling, system builds and records each speaker model that presented by feature vectors extracted in front-end processing steps. Many modeling techniques now are introduced in speaker recognition, and the selection of modeling is not only dependent on the type of speech to be used, desired performance, but also the ease of training and updating, and storage and computation considerations. Then what does a speaker model like? Douglas A. Reynolds presents it in his paper as follows: “Desirable attributes of a speaker model are: (1) a theoretical underpinning so one can understand model behavior and mathematically approach extensions and improvements; (2) generalizable to new data so that the model does not over fit the enrollment data and can match new data; (3) parsimonious representation in both size and computation.”[1]

2.3. Pattern Matching

The pattern-matching task of speaker recognition involves computing a match score or distance between two speaker models, and the score or the distance can measure the similarity to the feature vectors each model built in Speaker Modeling step. After enrolling users into the system, then the matching algorithm computes the distance between the user and each of speaker model, and identify the speaker by the maximum score (minimum distance), or compares/scores the incoming speech signal with the model of the claimed user. Pattern-matching methods includes models such as dynamic time warping, the hidden Markov model and vector quantization, responding to template models, statistical models and codebook models.[3]

3. Voice over IP (VoIP)

VoIP refers to a way to carry phone calls over an IP data network, whether on the Internet or your own internal network. A primary attraction of VoIP is its ability to help reduce expenses because telephone calls travel over the data network rather than the phone company’s network. VoIP and IP telephony are becoming increasingly popular with large corporations to consumers alike. Internet Protocol (IP) is increasingly viewed as more than just a way to transport data, but also as a tool that simplifies and streamlines a wide range of business applications. VoIP (voice over IP) is also the foundation for more advanced unified communications applications that can have an increasingly transformative effect on the way people communicate.
3.1. VoIP Protocol Stack

As its name implies, VoIP utilizes IP as its basic transport method. There are a number of protocols that may be employed in order to provide the VoIP communication services. VoIP utilizes both the TCP and UDP protocols over IP. It is important to note that VoIP works with any protocol stack that supports IP. End users of VoIP can add enterprise VoIP systems to their existing infrastructure relatively quickly and easily. The Voice over IP (VoIP) protocol suite is generically broken into two categories, control plane protocols and data plane protocols. The control plane portion of the VoIP protocol is the traffic required to connect and maintain the actual user traffic. It is also responsible for maintaining overall network operation (router to router communications). The data plane (voice) portion of the VoIP protocol is the actual traffic that needs to get from one end to another.

3.2. Control Plane Protocols

The control plane is used for the various signaling protocols, allowing users of VoIP to create and maintain their phone calls. The most prevalent types of signaling protocols today are H.323 and SIP.

H.323 is ratified by the International Telecommunication Union-Telecommunication (ITU-T). It is designed to operate above the transport layer of the underlying network, so H.323 can be used on TCP/IP which is packet-based network transport. H.323 specified protocols including Q.931, H.225, H.245 and ASN.1.

Session Initiation Protocol (SIP) is designed to manage and establish multimedia sessions, such as video conferencing, voice calls, and data sharing. SIP is specified as proposed standard RFC 3261. This is the standard that many element manufacturers are using to develop products. SIP is a text-based protocol, similar to HTTP and SMTP, for initiating interactive communication sessions between users.

H.323 is more complex than SIP, making it harder for network managers to troubleshoot and more complex for application developers to work with. SIP is more in the next-generation space, while H.323 is more of a bridge between older and newer approaches. And this paper focuses on SIP.

3.3. Data Plane Protocol and Codec

Once signaling and encoding are ready, Real-Time Control Protocol (RTCP) and Real-Time Control Protocol (RTCP) are used to deliver voice packets. RTP is a data plane protocol that is used infrequently. This protocol allows the endpoints to communicate directly regarding the quality of the call. RTCP affords the endpoints the ability to adjust the call in real time to increase the quality of the call. RTCP also aids significantly in the troubleshooting of a voice stream. Traditional VoIP analyzers sit at specific locations on a circuit and base their derived results from only the packets that they capture. With RTCP enabled, the analyzer can see the end-to-end quality as well as the quality at the point at which the analyzer is inserted.

There is a wide range of voice Codec (coder/decoder or compression/decompression) protocols available for VoIP phone implementation. The most common voice Codec includes G.711, G.723, G.726, G.728, and G.729, etc.

4. Experiments

4.1. Applying Speaker Recognition on VoIP Auditing

After having knowledge of Speaker Recognition and VoIP’s communicating principle, now we can apply speaker recognition on VoIP auditing. It can be computed as follows:

1. Speaker modeling in speaker recognition
2. SIP protocol analysis in VoIP auditing
3. RTP protocol analysis in VoIP auditing
4. Pattern-matching in speaker recognition

The first step is the same as speaker modeling in common speaker recognition system and the other steps will be described right next.

4.2. VoIP Auditing

As explained before, VoIP includes two main kinds of communication: Signaling protocol communication and Media data (RTP) communication, so VoIP auditing is divided in two parts:

1. Signaling protocol analysis
2. Voice data traffic analysis

Before capture voice data, we should analyze signaling
protocol first. Because from the traffic monitoring point of view, there are some reasons listed as follows:

- The signaling protocol contains important information such as parties’ identity, type of call, codec, duration, and information about the RTP sessions(s) that usually do not take place on fixed ports. In general without properly decoding the signaling protocol, RTP connect cannot be detected exactly, and it is not possible to guess the ports used for RTP.

- Beside voice, the transport protocol also can be used to extract information such as jitter, packet loss, and packet latency that are the building blocks for evaluating the call quality. Note that as the RTP packets contain information about the codec being used, with appropriate software it is also possible to decode the RTP and extract further call information.

And the first reasons are the key point of this application for VoIP auditing. Furthermore as RTP traffic is flowing on dynamic ports, packet filtering facilities provided by standard equipment are not suitable as they are static and not able to be reconfigured on the fly based on the signaling protocol.

In this application, we choose SJPhone to build VoIP communication environment. SJPhone is a VoIP soft phone that allows you to speak with any other soft phone because it utilizes standard VoIP protocol. And we configure it with setting SIP as its signaling protocol. According SIP protocol and message format, we can obtain its data and get SIP information when signaling procedure by listening its port.

The body of a SIP message is defined by the Session Description Protocol (SDP). The Session Description Protocol (SDP) is an IETF standard for announcing and describing multimedia conferences. A session description consists of three parts: a single session description, zero or more time descriptions, and zero or more media descriptions. The session description contains global attributes that apply to the whole conference or all media streams. Time descriptions contain conference start, stop, and repeat time information. Media descriptions contain details about a particular media stream.

Figure 4. SIP Protocol Packets

![Figure 4. SIP Protocol Packets](image)

Figure 5. Media descriptions in SIP Message Body

The most important that we need is the Media descriptions. From this message body, RTP port and coder algorithm is negotiated. In the figure above, Media descriptions indicate that the RTP port is 49156 and the payload type it supports are 0 and 101. In RTP, the coder PCMU is used for type 0 and dynamic for type 101, and the former has priority.

Now signaling protocol analysis is finished, and voice data traffic analysis is started. As known before, RTP port is obtained when analysis SIP protocol, then we should detect RTP connection first. So we should adjust the sniffer port setting from SIP port to RTP port.

Figure 6. RTP Protocol Packets

![Figure 6. RTP Protocol Packets](image)

Figure 7. Flow of VoIP Auditing

![Figure 7. Flow of VoIP Auditing](image)
4.3. Identify Speaker by Speaker Recognition

Once we have decoded the RTP packets into wave files, the input data for speaker recognition are ready. In our experiment, we build up a speech corpus with 10 boys and 5 girls while VoIP communication. This Speaker Recognition Module selects Mel-Warped Cepstrum and combining DTW and VQ to build and recognize speaker model.

![Speaker Recognition under Different Codec](image)

Figure 8. Speaker Recognition under Different Codecs

From the result we can see three different recognition rates under three different codec. Voice must be encoded for transmission and subsequently decoded at the far end. Furthermore, the compression rate is usually high and consequently, the encoding-decoding process causes an impoverishment of the recognition figures. When speech data is compressed by G.711 u law, the recognition rate is highest, and lowest by G.729 compressed. Because G.711 converts voice into a 64 kbps voice stream. This is the same codec used in traditional TDM T1 voice. It is considered the highest quality, while G.729 converts voice into an 8 kbps stream. Voice will lose more information when compressed by G.729 compared to G.729.

Moreover, because VoIP’s voice data is transferred over internet, IP networks are not intended for transmitting voice. This unsuitability constitutes a very challenging problem such as packet loss, delay, and network jitter. These will reduce the recognition performance.

5. Conclusion and Future Work

Though Speaker Recognition has been researched for many years, along with spread use of VoIP and more and more rampant crime over Internet, combining VoIP auditing and Speaker Recognition can be an effective approach to reduce and investigate this kind of crime. This paper proposes and presents a mechanism of capturing voice data and identifies the responding speaker. VoIP auditing includes signaling protocol analysis and voice data traffic analysis, and the latter is based upon the former. Promising result shows that speaker recognition is effective, but its performance is reduced by two obstacles: lose information of codec and packet loss.

In the near future, one question of this system should be considered is how to analyze SIP protocol exactly and capture RTP payload in real-time. So integration of VoIP voice data capture and speaker recognition should improve robustness and real-time responding. And to recognize the speaker from the compressed voice data is a direction of future research.

Acknowledgements

I would like to thank my supervisor, and my classmates of ICRC. I have received great assistance from them. To everyone who has contributed, I owe my deepest appreciation.

References