Reviewing Speaker Recognition Influence in Strengthening VOIP Caller Authentication

Qutaiba Ali, Nada Abdul Ghani
Computer Engineering Department, University of Mosul, Iraq

Abstract*: This paper focuses on authentication issues and user ID detection depending on the voice signature of the incoming Voice over IP (VoIP) calls. Firstly, a complete speaker recognition system based on LPC feature extraction and decision based on Artificial Neural Network was described and built using MATLAB package. Then, MATLAB package was used to develop a simplified prototype version of the modified SIP protocol which is used to manage calls establishment among the network nodes. These codes were written for both SIP server and client nodes. The modified SIP protocol was implemented on an experimental network which is built to demonstrate the operation of the enhanced authentication procedure.*

Keywords*: Voice over IP, Session Initiation Protocol, Speaker Recognition, Neural Network.*

Received August 30, 2009; Accepted February 17, 2010

1. Voice over IP (VoIP) Security

Voice over IP, or Internet telephony is the idea is to use the Internet as a telephone network with some additional capabilities. Instead of communicating over a circuit-switched network, this application allows communication between two parties over the packet-switched Internet.. The VoIP data processing consists of the following four steps: signaling, encoding, transport, and gateway control [11, 12].

- Signaling: The purpose of the signaling protocol is to create and manage connections or calls between endpoints.H.323 and the Session Initiation Protocol (SIP) are two widely used signaling standards for call setup and management. SIP has proven to be the most widely used signaling protocol, and hence it is considered here in this work [23, 24].
- Encoding and Transport: Once a connection is setup, voice must be transmitted by converting the voice into digitized form, then segmenting the voice signal into a stream of packets. The first step in this process is converting analog voice signals to digital, using an analog-to-digital converter. Here a compression algorithm can be used to reduce the volume of data to be transmitted. Next, voice samples are inserted into data packets to be carried on the Internet using typically the Real-time Transport Protocol (RTP). RTP packets have header fields that hold data needed to correctly reassemble the packets into a voice signal on the other end. Lastly, the encapsulated voice packets are carried as payload by the user datagram protocol (UDP) for ordinary data transmission. At the other end, the process is reversed: the packets are disassembled and put into the proper order, and then the digitized voice is processed by a digital-to-analog converter to render it into analog signals for the called party's handset speaker. There are 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 [11l 12].
- Gateway Control: The IP network itself must then ensure that the real-time conversation is transported across the telephony system to be converted by a gateway to another format—either for interoperation with a different IP-based multimedia scheme or because the call is being placed onto the PSTN [23, 24].

In fact, the Internet is the foundational service used by Voice over IP (VoIP), and the Internet itself can be considered a hostile environment from the security point of view. All Internet users must protect their transmissions from potential attackers, and VoIP users are no exception. This is true for both security of SIP-enabled sessions and also for SIP signaling security [21].

Authentication is a security feature which ensures that access is given only to users who are permitted access. During a call involving SIP user-agent and server, an attacker could masquerade as a user, forging the real identity of the sender. Authentication provides a mechanism to verify that a request sender and/or receiver are legitimate. Each phone has an identity (Usually phone number) that is associated with the device. Having a device impersonate the identity of another can be used as an attack to either receive calls or place calls with the spoofed identity. An attacker wishing to impersonate someone would setup their VOIP phone device to use the identity of the victim's device. The attacker's device then registers with the phone system. Any calls intended for the victim's phone number would then be directed to the attacker's phone. The attacker could then answer a call and impersonate the victim. The attacker also has the ability to use the spoofed device to place calls. The caller ID at the phone receiving the call would indicate that the victim is calling, not the attacker. This can also be used by the attacker to impersonate the victim [10].

Another attack is IP Spoofing, an attacker gains unauthorized access to a computer or a network by making it appear that a malicious message has come from a trusted machine by "spoofing" the IP address of that machine. This technique can be used to impersonate either a control node or an end node in the VOIP network [20].

Caller ID information has become relied upon for many purposes. Many banking applications now use caller ID to verify the identity of the customer. An attacker spoofing the ID of a victim's phone could use this to gain access to banking or credit card information [7].

The traditional phone system is minimally susceptible to this type of attack. The centralized configuration of phone lines in the switched telephone network prevents the spoofing of caller ID information for most lines. This general acceptance that the call ID can be trusted has led to the development of applications that rely on caller ID to establish the identity of the person. This same level of integrity needs to be maintained in the VOIP-based phone system to be accepted by the general public [20].

The comprehensive overview on VoIP security is the reference [7] addressing the operational and deployment aspects of VoIP security. The security mechanisms deployed in SIP are well described in [8] without covering the formal aspect of the security architecture.

Many works have been dedicated to analysis and testing of VoIP protocols, but dealing either with the PSTN interconnection as in [17], or [18]. Most of the performed work has addressed the prevention of SPAM over Internet Telephony (SPIT) attacks [19] as well as mitigating denial of service ones (DOS) [22].

Several frameworks have been proposed to address specific aspects of VOIP security. The SIP Forum Test Framework [1] is a conformance test suite that allows SIP device vendors to test their devices for common protocol errors and thus improve inter-operability. The SNOCER project [2], [6] proposes a generic framework to protect SIP from malformed message attacks and describes a signature based detection mechanism for SIP message tampering.

This paper will focus on authentication issues and user ID detection depending on the voice signature of the incoming call. We suggest the use of Speaker Recognition (SR) concepts to support user identification process.

2. Proposed Speaker Identification System

Speaker recognition is a generic term used for two related problems: speaker *identification* and *verification*. In the identification task the goal is to recognize the unknown speaker from a set of *N* known speakers. In verification, an identity claim (e.g., a username) is given to the recognizer and the goal is to accept or reject the given identity claim. This paper concentrates on the identification task [3].

The input of a speaker identification system is a sampled speech data, and the output is the index of the identified speaker. There are three important components in a speaker recognition system: the feature extraction component, the speaker models and the matching algorithm. Feature extractor derives a set of speaker- specific vectors from the input signal. Speaker model is then generated from these vectors for each speaker. The matching procedure performs the comparison of the speaker models [25].

Speaker identification system can be either text-dependent or text-independent. The current work makes use of the text-dependent speaker identification systems which requires that the speaker should utter the same phrases as that use in system training session. On the other hand, text-independent speaker identification systems identify the speaker regardless the text of his utterance. The text of speaker's voice could be either in fixed vocabulary or free both in training and testing sessions. In the case of text free, the system needs much more training and testing speech data than that in the case of text confined in fixed vocabulary [15].

Speaker identification system can conclusively be modeled as shown in Figure (1). Features are some quantities, which are extracted from preprocessed speech and can be used to represent the whole speech signal.

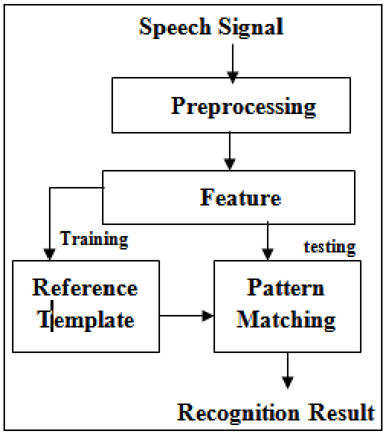


Figure 1. Speaker recognition model.

Two sets of features consisted of spectral and prosodic features are usually used. However, spectral based features such as Linear Prediction Coefficients (LPC), Cepstral coefficients, and their derivatives have proven to be more efficient than prosodic-based features like fundamental frequency, formant frequency, and speech energy [3, 25, and 15].

There are many recognition system proposed for pattern matching (speaker recognition) stage such as a well known nonlinear time-aligned techniques called Dynamic Time warping (DTW), Vector Quantization (VQ), Artificial Neural Network (ANN), and statistical Hidden Markov Model (HMM) [14, 5] . In the current work, Backpropagation Artificial Neural Network (BANN) were used for pattern matching purposes.

In our proposed system, digital speech signal is passed through a preprocessing procedure, which performs energy-based end point detection and time normalization. LPC features are then extracted from preprocessed speech and the input vectors have passed through the backpropagation learning algorithm with multilayered perceptron network for both training and evaluation processes [16]. Figure 2 shows overall model of the proposed system.

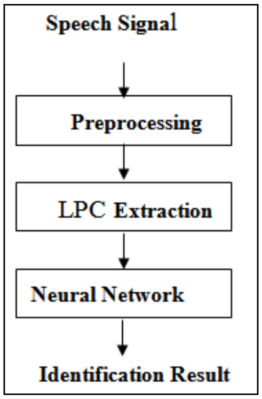


Figure 2. Overall proposed speaker identification model

The following sections summarize the design principles of the system [4-9].

2.1. Preprocessing Stage

The function of this stage is to prepare the voice samples to LPC extraction phase and has two actions: Pre-emphasis and Framing [4, 9].

2.2. Feature Extraction

Next, we describe the procedure for computing the feature vectors from a given speech signal *s(n)*. The most commonly used features in speaker recognition systems are the features derived from the cepstrum [4], [9]. In our system, the number of LP coefficients (dimension of feature vectors) was selected as 14, which are the input to linear network.

2.3. Multilayer Neural Network

In this paper, we chose to use backpropagation neural networks see figure 3 since it has been successfully applied to many pattern classification problems including speaker recognition systems [13]. In our system, the input layer contains 14th neuron which depends upon the coefficient of LPC, hidden layer depends upon the experimental work (40 neurons) and output layer contains four neurons depending upon number of the recognized speakers.

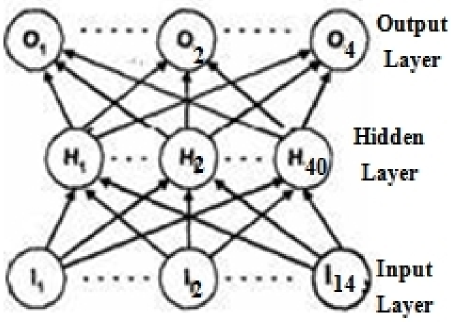


Figure 3. Backpropagation neural networks.

3. Analysis of the Suggested VOIP Security Method

Our approach suggests that any caller to a VoIP entity must proves his identity by giving an online spoken password to a VoIP security server before call establishment process could be completed. The VoIP security server consists of a speaker recognition engine and a database of spoken passwords to the legal system users. Any user wishes to subscribe to this VoIP system should give a password spoken by him which is then stored in the database mentioned earlier. In such a manner, any caller is authenticated by comparing his spoken password against the stored ones. This method enhances user ID verification by adding voice signature as a part of user identification procedure.

As shown in igure 4, the traditional VoIP system could be enhanced by the addition of Speaker Recognition server coupled with SIP server of the system.

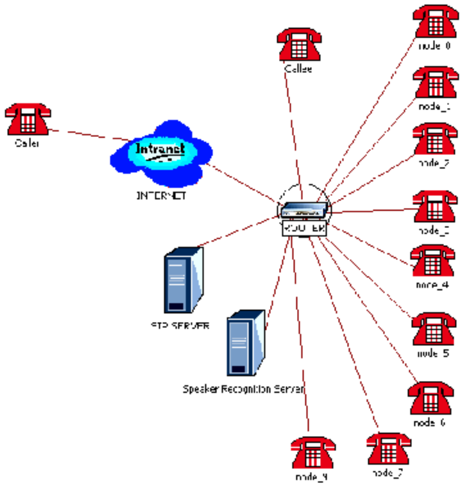


Figure 4. The Suggested VOIP System.

The operation of the traditional VOIP system should be modified according to the following procedure:

- Any user wishes to register to this VOIP system is asked by SIP server to provide a Spoken Password. The server must supply the client with a temporary key (using a certain key management and distribution protocol) to encrypt this spoken password to prevent eavesdropping during this phase. The user pronounce his password which is then encrypted in real time and passed to the Speaker Recognition server to store it in its database (actually, this is equivalent to the learning phase in Neural Network operation).
- According to a previous agreed criteria between them, client and server sides will extract a symmetric encryption key from this spoken password. We suggest the following simple procedure to achieve this task:
  - Using an equivalent and synchronized Pseudorandom Number Generator program in both sides (with the same seed value), a certain number is generated. The length of this number is (3N of bits), where N represents the ciphering key length.
  - Choosing three sections from the beginning, middle and end of the spoken password binary file, each has N bits length.
  - X-OR operation is performed between the two patterns, then, N bits checksum is calculated for the result. The resultant N bits number is the required ciphering key.
- During call establishment procedure, SIP server will ask the caller to prove his identity by giving the same spoken password given earlier.
- The client pronounces his password which in then encrypted together with its hash function and sent to SIP server. On the reception of this combination, SIP server decrypt the incoming message and recomputed the hash function to insure its integrity, see Figure(5). The above procedure must be followed to eliminate the possibility of intrusions and attacks during this phase and to authenticate the incoming message.
- The given spoken password is forwarded by SIP server to the speaker recognition server to compare it against previously stored passwords.
- Speaker recognition server performs its checking function and returns the result to SIP server. Depending on this result, SIP server may (or may not) allows finishing call establishment procedure. This method would enhance user ID detection because it depends on the vocal properties of caller in addition to the password value. The integration between these parameters supports Authentication process and gives immunity against previously mentioned attacks.

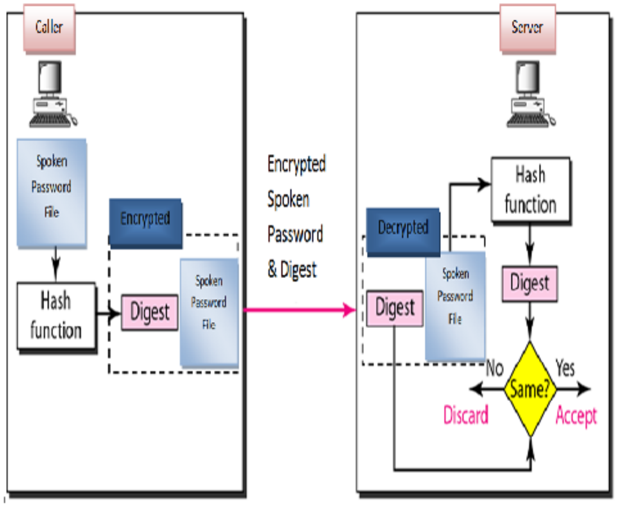


Figure 5. Message Authentication Procedure.

4. Experimental Setup of the Method

In order to test the affectivity of the suggested method, an experimental network was built. In this experiment, the utterance is obtained from 4 local speakers (2 females and 2males). Each speaker was asked, in an office environment, to pronounce his/her password which is a combination of the digits from zero to nine in Arabic language. These phrases were used to train the neural network of the speaker recognition system.

The average length of password sentence was 5 second. In this experiment, G.711 Codec was used in which voice samples are encoded at 64kbps without compression.

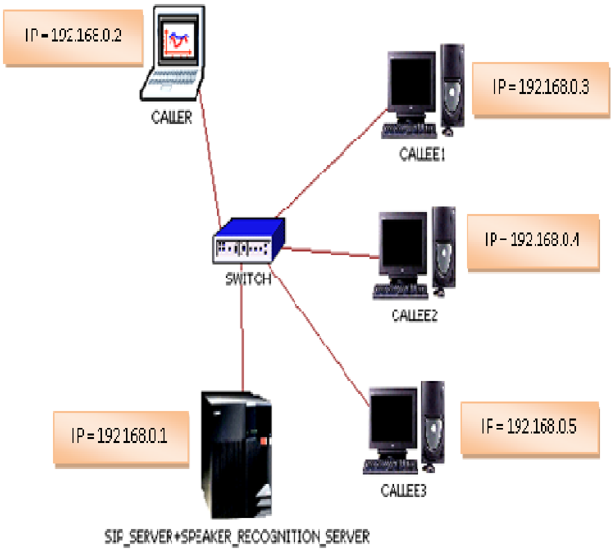
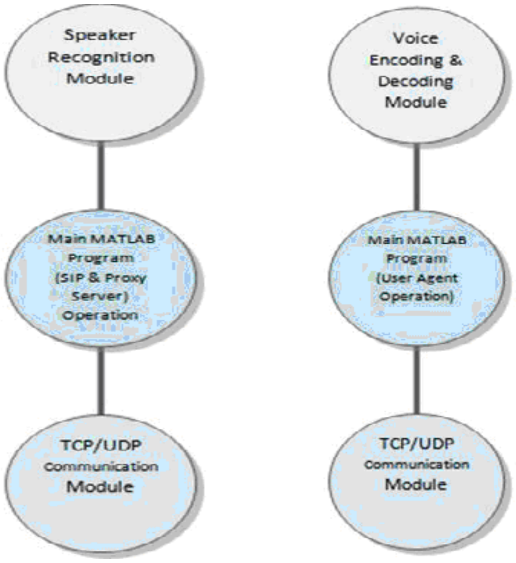


Figure 6. The Experimental Testbed.

The experimental network consists of the following , see Figure 6:

- Fast Ethernet LAN (100 Mbps data rate).

- Four personal computers to represent the user agents of VOIP service (2GHZ Core2Due Processor, 1GByte RAM).
- For the sake of simplicity, speaker recognition server and SIP server were implemented on the same platform (2GHZ Core2Due Processor, 2GByte RAM). Also, this server was supplied with some *Proxy Server* features (i.e., managing network access rules), in which requests and responses from user agents initially are made through the Proxy server.
- MATLAB package was used to develop a simplified prototype version of the modified SIP protocol which is used to manage calls establishment among the network nodes. As shown in Figure (7), MATLAB codes written for SIP server consists of the main program used to implement SIP and Proxy server and two modules engaged with it : Speaker Recognition and TCP/UDP communication modules. On the other hand, client side consists also of three parts: main program, voice encoding/decoding module and TCP/UDP communication module.


Speaker Recognition Module
Main MATLAB Program (SIP & Proxy Server) Operation
TCP/UDP Communication Module
Voice Encoding & Decoding Module
Main MATLAB Program (User Agent Operation)
TCP/UDP Communication Module

(a)Server Side. (b) Client Side.

Figure 7. Structure of MATLAB Codes.

The experiment flows were as follows, see Figure 8:

- In this experiment, IP addresses were used to act as phone numbers in an VOIP system. The caller must give the IP address of the called party in order to get connected with it.
- One of the nodes were supposed to be the Caller who is attempting to invite one of the Callees into a session. He begins by sending a request (initiating a TCP/IP session) to the SIP-Proxy server.
- The caller is asked to prove his identity by supplying the server with an online spoken password.
- On the reception of this challenge (a number (4455) which represent a prior agreed command), MATLAB program in the caller side asked the user to pronounce his password via microphone device. The digitized voice samples is converted into a binary file and the Hash function of this file is computed (a Ready to use MATLAB code for SHA-1 hash function was imported for this purpose). The spoken password file together with its hash function is encrypted using Data Encryption Standard (DES) (a Ready to use MATLAB code for DES method was imported for this purpose) and packetized (in one or more Ethernet packets) and sent to the server.
- The server decrypts and authenticates the received packets then checks the received spoken password with the speaker recognition module to determine the ID characteristics of the caller.
- Building on the caller ID investigation, the SIP-Proxy server decides whether to accept or deny the call request. If the result was negative (unauthorized caller), the server terminate the connection. Otherwise, the server passes the connection request to the called party and the caller issued to start his conversation using UDP protocol.

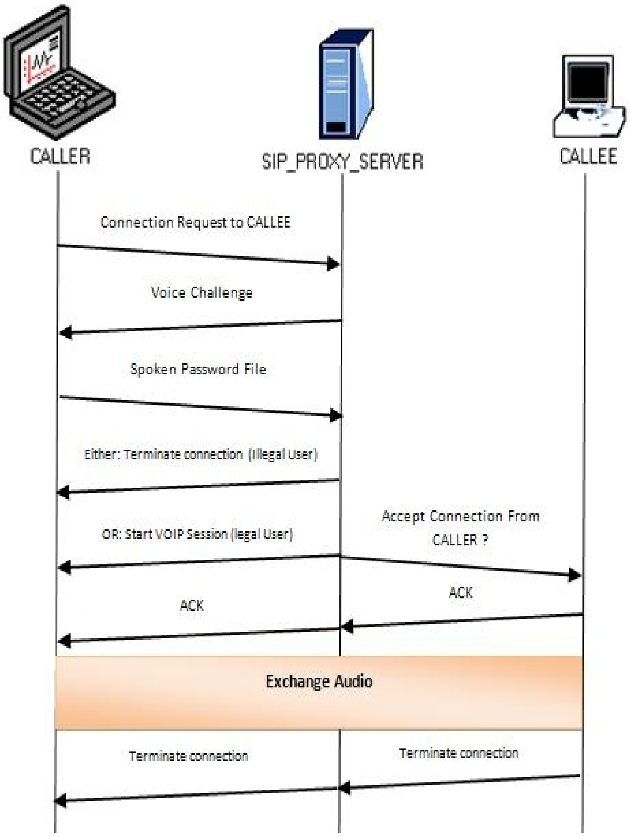

CALLER
SIP_PROXY_SERVER
CALLEE
Connection Request to CALLEE
Voice Challenge
Spoken Password File
Either: Terminate connection (Illegal User)
OR: Start VOIP Session (legal User)
Accept Connection From CALLER ?
ACK
ACK
Exchange Audio
Terminate connection
Terminate connection

Figure 8. Flows of the Experimental Network

The former experiment repeated four times to cover the different possibilities as follows:

- Legal registered user with (correctly spoken- true password), the result was “Access Granted” with 100% speaker identification rate.
- Legal registered user with (correctly spoken- false password), the result was “Access Denied” with 90% speaker identification rate.
- Illegal user with (incorrectly spoken- true password), the result was “Access Denied” with 0% speaker identification rate.
- Illegal user with (incorrectly spoken- false password), the result was “Access Denied” with 0% speaker identification rate.

5. Conclusions and Suggestion for Future Work

This paper suggests the adoption of speaker recognition methods to enhance user ID detection of VOIP system. This method would enhance user ID detection because it depends on the vocal properties of caller in addition to the password value. The integration between these parameters supports authentication process and gives immunity against caller ID spoofing attacks.

References

[1] “SIP Forum Test Framework (SFTF),” http://www.sipfoundry.org/sipforum-test-framework/.

[2] “SNOCER: Low Cost Tools for Secure and Highly Available VoIP Communication Services,” http://www.snocer.org/.

[3] Campbell J.P., “Speaker Recognition: A tutorial,” Proc. IEEE, Vol.85, pp.1437-1462, Sept.1997.

[4] Chen Y., “Cepstral Domain Talker Stress Compensation for Robust Speech Recognition”, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 36, no. 4, pp. 433-439, Apr. 1988.

[5] Deller J., Proakis J.G., and Hansen J., “Discrete-Time Processing of Speech Signals”, Macmillian Publishing, 1993.

[6] Geneiatakis D., Kambourakis D., “SIP Message Tampering: THE SQL code INJECTION attack,” in Proc of 13th Intl Conf on Software, Telecoms and Comp Networks IEEE, Split, Croatia, Sept, 2005.

[7] Geneiatakis D., Dagiuklas T., Kambourakis G., Lambrinoudakis C., Gritzalis S., Ehlert K., and Sisalem D., “Survey of security vulnerabilities in session initiation protocol. Communications Surveys & Tutorials”, IEEE, 8(3):68–81, 3rd. Qtr. 2006.

[8] Johnston A.B. and Piscitello D. M., “Understanding Voice over Ip Security (Artech House Telecommunications Library)”, Artech House, Inc., Norwood, MA, USA, 2006.

[9] Lee C.H., “On Robust Linear Prediction of Speech”, IEEE Transactions on Acoustics, Speech, and Signal Processing”, vol. 36, no. 5, pp. 642-650, May 1988.

[10] McGann S. and Sicker D., “An analysis of security threats and tools in SIP-based VoIP systems”, presented at the 2nd Annu. Workshop VoIP Secur.,. Washington, DC, Jun. 2005.

[11] Metha P. and Ubani S., “Voice over IP,” IEEE Potentials Magazine, Vol. 20, Oct. 2001.

[12] Mehta P. and Udani S., “Overview of voice over IP”, Dept. Computer. Inf. Sci., Univ. Pennsylvania, Philadelphia, PA, Rep. MS-CIS-01-31,Feb. 2001.

[13] Oglesby J. and Mason J.S., “Optimization of neural models for speaker identification”, in Proc. IEEE Int. Conf. Acoustics,Speech, Signal Processing (ICASSP ’90), vol. 1, pp. 261–264, Albuquerque, NM, USA, April 1990.

[14] Picone J., “Signal Modeling Techniques in Speech Recognition,” Proceedings of the IEEE, vol. 81, no. 9, pp. 1215-1246, Sept. 1993.

[15] Quatieri T.F., “Discrete-time Speech Signal Processing Principles and Practice”, Prentice-Hall, Inc., New Jersey, 2002.

[16] Rabiner L. and Juang B.H., “Fundamentals of Speech Recognition”, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1993.

[17] Sengar H., Dantu R., and Wijesekera D., “Securing voip and PSTN from integrated signaling network vulnerabilities”, 1st IEEE Workshop on, pages 1–7, 3 April 2006.

[18] Sengar H., Dantu R., Wijesekera D. and JajodiaS., “SS7 over IP: Signaling internetworking vulnerabilities”, IEEE Network, Vol. 20, No. 6, pages 32–41, November, 2006.

[19] Sisalem D., Kuthan J. and Ehlert S., “Denial of service attacks targeting a sip voip infrastructure: attack scenarios and prevention mechanisms”, IEEE Network, 20(5):26–31, Sept.-Oct. 2006.

[20] Sicker D. and Lookabaugh T., “VoIP security: Not an afterthought,” ACM Queue, vol. 2, no. 6, pp. 56–64, Sep. 2004.

[21] Stallings W., “Cryptography and Network Security”, Upper Saddle River, NJ: Prentice-Hall, 2003.

[22] Thermos P. and Takanen A., “Securing VoIP Networks: Threats, Vulnerabilities, and Countermeasures”, Addison-Wesley Professional, 2007.

[23] Varshney U., “Voice over IP”, Communications of the ACM, Vol.45, No. 1, Jan. 2002.

[24] Walsh T.J. and Kuhn D.R., “Challenges in securing voice over IP”, IEEE Security & Privacy Magazine, Vol. 3, Issue 3, May 2005.

Zilca D., “Text-Independent Speaker Verification Using Covariance Modeling”, IEEE Signal Processing Letters Vol.8.No4, April 2001.

Qutaiba Ali was born in Mosul, Iraq, on October ,1974. He received the B.S. and M.S. degrees from the Department of Electrical Engineering, University of Mosul, Iraq, in 1996 and 1999, respectively. He received his Ph.D. degree (with honor) from the Computer Engineering Department, University of Mosul, Iraq, in 2006. Since 2000, he has been with the Department of Computer Engineering, Mosul University, Mosul, Iraq, where he is currently an assistance professor. His research interests include computer networks analysis and design, embedded network devices and network security. Dr. Ali instructed many topics (for Post and Undergraduate stage) in computer engineering field during the last ten years and has many publications in numerous journals and conferences. He acquires many awards (form National Instrument INC.) and appreciations form different parties for excellent teaching and extra scientific research efforts. Also, he was invited to join many respectable scientific organizations such as IEEE, IENG ASTF, WASET and many others. He was participate (as technical committee member) in three IEEE conferences in USA, Malaysia and Egypt and joined the editorial board of five scientific international journals.

Nada Abdul Ghani was born in Mosul, Iraq, on 1955. She received the B.S. from the Department of Electrical Engineering, University of Mosul, Iraq, in 1977. She received her MS degree from the Computer Engineering Department, University of Mosul, Iraq, in 2004. Since 1977, he has been with the Department of Electrical Engineering, Mosul University. In 1999, she joined computer engineering department, Mosul University, where she is currently an assistance Lecturer. Her research interests include Voice & Speaker Recognition using artificial Intelligence techniques.