A Novel Watermarking-based Integrity Method for Internet Telephony in WLAN Environments

Sorin A. Huss and Song Yuan
ISS, Department of Computer Science
Technical University of Darmstadt
Hochschulstrasse 10, Darmstadt
64289, Germany

Abstract: - Data integrity and source origin authentication are essential topics for VoIP systems in wireless LAN. But traditional methods, such as MAC, are not well-suited to overcome the high distortion rate introduced by audio data transportation over wireless LANs. In this paper a new integrity mechanism deploying speech watermarking robust to the distortion in wireless LANs is presented. In addition, the advocated approach adopts public key encryption to efficiently generate non-repudiate speech.Finally, we propose a new speech watermarking algorithm incorporated with the GSM 610 full-rate coder.

Key-Words: - Speech Watermarking, Integrity, VoIP, Wireless LAN

1 Introduction

VoIP or Internet telephony is the transportation of voice traffic over the Internet protocol (IP). The Internet acts an interconnection between networks that use IP. Over the years the Internet has become a basis for new applications and services it was not intended for at the beginning. It grew to a market with high potentials especially for services such as IP telephony. However, lot of security incidents occurred in the last years and their number is still increasing. Therefore, security is an essential topic for VoIP telephony systems [2].

There are five different categories of security services present for IP telephony: Identification and authentication, Authorization, Confidentiality, Integrity, Non-denial/Non-repudiation. These security aspects can be provided by means of different cryptographic techniques such as secret key cryptography, public key cryptography, and hash functions, respectively. In this article, we mainly address integrity and source origin authentication of multimedia data. By multimedia data integrity and authenticity mechanisms, we can provide some very useful related information, such as the speaker identity, the point in time when dialog happened, and the integrity/authenticity of the transaction.

At present IEEE 802.11 wireless LANs [1] are spreading very fast. Applying Internet Telephony technique over WLANs could be a low-cost voice communication means within enterprise or campus networks. In this paper we investigate a new digital watermarking-based integrity method for VoIP implementation over WLAN which is robust to considerable distortions often occurring in wireless networks. The scenario we are considering is shown in Fig. 1. The network comprises a single IEEE 802.11 basic service set (BSS) with one access point (AP) and a number of mobile users. The AP is connected to a 100 Mb/s Ethernet, to which other users are directly connected. Voice calls take place between an user in the BSS and an user connected to the Ethernet (i.e., between the users `Sender` and `Receiver1`).

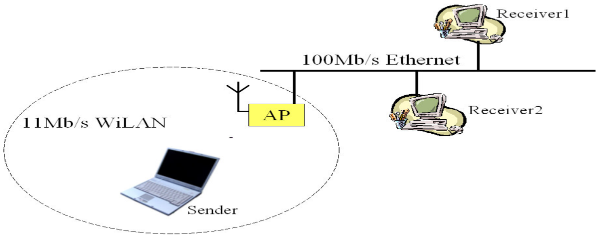

Fig. 1 Network Scenario

Internet telephony unveils some special properties stemming from both multimedia data and real-time communication. First, end users cannot perceive limited distortions in multimedia data, so, some bit errors and packets losses occurring during communication do not defect the overall visual/audio quality. Secondly, due to controlling protocols and implementations of real-time multimedia communication, packet losses may happen any time. Thirdly, caused by the large amount of multimedia data, the communication security trade-offs should be as low as possible.

1.1 Secure Real-Time Transport Protocol

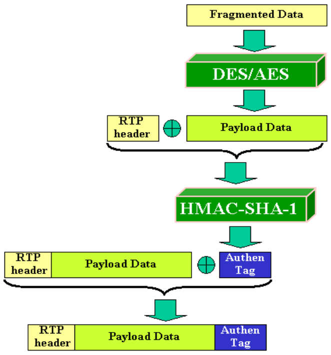

Fig.2 Integrity and Confidentiality Mechanism of SRTP

SRTP [3] is the IETF recommended standard for data integrity of Internet Telephony. SRTP supports both message integrity protection and source origin authentication. For integrity protection, a message authentication tag is appended to the tail of the RTP packet as depicted in Fig. 2. This tag is ranging over the entire RTP packet and it is computed after the packet has been encrypted. The HMAC-SHA-1 integrity protection algorithm is specified for use with SRTP. The method can be expressed as:

1) **Data Encryption**

$$Payload\_Data_{encrypted} = DES(\ key_{DES},\ Payload\_Data\ )$$

2) **Hashing**

$$Authen\_Tag = SHA(\ Payload\_Data_{encrypted}\ )$$

Source origin authentication using the TESLA (Timed Efficient Stream Loss-tolerant Authentication) algorithm has been suggested for this purpose.

1.2 Drawbacks of SRTP

MAC was originally developed as a solution to the integrity and authenticity problem for text messages. So, it does not take into consideration explicitly the special properties of multimedia data nor of real-time communication. Thus, there are some serious drawbacks when one attempts to exploit these methods for the envisaged application.

First, MAC is sensitive to data distortion — an unrecoverable bit error in the multimedia message or in the digital signature/MAC code may disable the corresponding authentication procedures. Secondly, MAC is in general sort of a checksum of the message. Hence, it is apart from the message itself. A possible attacker may get access to the original multimedia message only and then reuse this message again and again to cheat the access control systems. Thirdly, these techniques increase the latency of multimedia communication so that the handheld devices (such as PDAs and cellular phones) in general cannot meet the resulting computing power requirements [4].

1.3 Fragile Digital Watermarking

Digital watermarking links some useful information to the multimedia data by embedding watermarks into the original data. In addition, an attacker cannot remove the embedded watermarks easily.

A large variety of audio watermarking algorithms has been proposed in the past and a few of them can be adapted for speech watermarking applications such as [5], [6], [7]. Many of these algorithms operate in time domain and exploit temporal or spectral masking models of the human auditory system. The most challenging requirement in real-time application, however, is to detect the fragile watermark in, say, every 0.5 seconds of speech with a reasonable error probability, especially when modern speech coders compress 0.5 sec of speech to less than 400 bytes of data.

The digital watermarks embedded in the original data may contain some useful information, e.g., author names, date of generation, or copyright holders. With the use of the blind digital watermarking algorithm, the embedded digital watermark can be extracted accurately without the need of the original multimedia data. Potential attackers, on the other hand, cannot retrieve the embedded watermarks without the knowledge of the private key or of the original data, respectively. Fragile audio watermarking algorithms can thus detect severe tampers/attacks occurred on the multimedia data. Therefore, it is a pretty useful method to ensure both authenticity and integrity of multimedia data.

Fragile digital watermarking provides an alternative approach to increase the safety of multimedia data during transmission in openly accessible channels. That is, digital watermarks may be generated referring to information on, say, originators, receivers, unique serial number, and time stamps. These watermarks are then embedded into the multimedia data to assure its integrity and origin source authentication without degrading the overall quality of the transmitted multimedia data.

2. An Integrity and Authenticity Mechanism for Real-Time Multimedia Communication

2.1 Integrity Model

SRTP protects the integrity of each data packet instead of the whole conversation data which consists of hundreds or thousands of audio data packets. In Internet telephony, an audio data packet contains the compressed sample for 20 ms, thus a packet loss does not affect the session too much. Fig. 3 illustrates a novel integrity model applicable for the whole session duration. Each speech data packet, p_k, has its own local integrity value, I_k, and the overall integrity denoted as I is defined as follows.

$$I = \frac{\sum_{k=1}^{N} I_k}{N} \qquad \textbf{(1)}$$

N represents the total number of the packets, and $\{ I_k \mid 0 \le I_k \le 1, 0 \le \mathbf{k} \le N\}$.

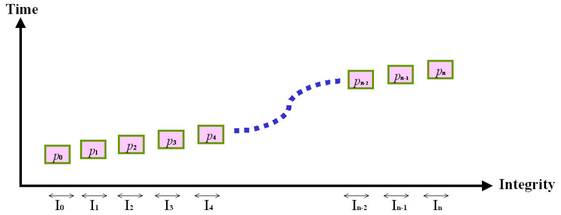

Fig.3 A New Integrity Model

Usually, the Internet telephone sends an RTP packet at an interval (denoted as ***t***) of 20ms or 40ms, and 0.5s of speech (denoted as ***s***) represents one accurate semantic meaning, so that the distortion of a RTP packet in a short period does not change the semantic meaning of the whole session. In other words, if a burst of distortion lasts for less than 0.25 s or for ***s***/2***t*** RTP packets being transmitted, then the quality of the session does degrade, but the integrity or authenticity of the multimedia communication is not destroyed. So, the proposed integrity and authenticity measurement can be taken on a range consisting of m (m << ***s***/2***t***) RTP packets to tolerate the high distortion rate of wireless LAN.

2.2 A Security Model Using Digital Watermarking Techniques

Digital watermarking provides an alternative approach to ensure the safety of multimedia data during transmission in openly accessible channels. The digital watermarks may be embedded into the multimedia data to assure its integrity and authenticity without degrading the overall quality of the transmitted data.

Fig. 4 shows the outline of a security mechanism using digital watermarks. Digital watermarking operates on the audio data to hide/extract information. This makes this approach different to most of the current cryptography mechanisms.

When inspecting the integrity and authenticity of the transmitted multimedia data, some special characteristics are to be taken into account - both the amount of the multimedia data and the occurrences of packet loss as well as of bit errors are unpredictable. Since the receiver can only receive the unpredictable multimedia data, we need an independent reference in order to verify the integrity and authenticity of the dynamic multimedia data during transmission. In the proposed approach, both the sender and receiver share one reference watermark (*comm_watermark*) from an independent trustworthy party before the multimedia data transmission really starts. To provide the non-repudiate and authenticity, a secret digital watermark, *speech_watermark*, is being introduced. *speech_watermark* has previously been encrypted with a public key algorithm, e.g., RSA.

The sender embeds both a public, *comm_watermark*, and a secret watermark, *speech_watermark*, into the outgoing multimedia data stream. The receiver then extracts the public digital watermark from its incoming stream. This scheme provides an integrated solution to secure multimedia communication and multimedia data integrity.

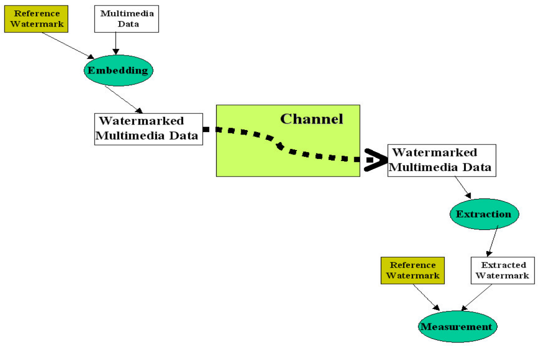

Fig. 4 Secure Real-time Communication with Digital Watermarking

The network transportation path can be viewed as a noisy channel. The reference digital watermark is modulated with multimedia data (carrier) and transmitted onto the noisy channel. The watermark undergoes the same changes suffered by the multimedia data, so that the watermark degradation can be used to estimate the overall alterations of the multimedia data caused by noise or by attacks. At the receiver side the embedded digital watermark is extracted and compared to the original reference watermark in order to measure the integrity and authenticity of the received multimedia data.

2.3 Outline of the Proposed Scheme

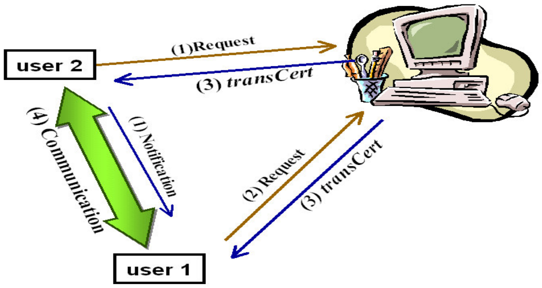

Fig. 5 Secure Multimedia Communication

At the beginning of each real-time multimedia communication, `user2` sends a *dialog_start* requests to *transServer* and notifies the callee, `user1`. Then, `user1` also sends a *dialog_start* requests to *transServer*. The *dialog_start* requests contain the digital certificates of the communicating parties. *transServer* authenticates the participants by verifying their digital certificates extracted from the received *dialog_start* requests. If acceptable, *transServer* generates a pair of *transCert*s, sends the *transCert*s to the participants, and stores them in its database, as depicted in Fig. 5.

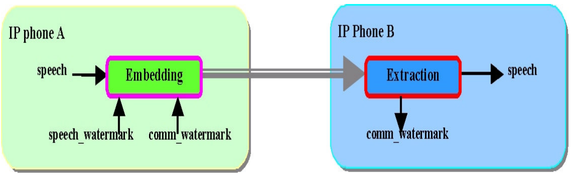

Fig. 6 Real-Time Authentication for VoIP

Once the participants receive the *transCert*s from *transServer*, the conversation may begin. At first, they authenticate the received transaction certificate. If the authentication is successful, then the participants parse the received *transCert*s to ((*speech_watermark, key*) *comm_watermark*) and insert then the digital watermarks into their outgoing audio data streams. At the same time, both participants extract the *comm_watermark* from their incoming streams and authenticate them, as depicted in Fig. 6. At the end of the transmission, i.e., conversation, each of the participants sends a *dialog_end* request to *transServer*. *transServer* authenticates the received *dialog_end* request and puts them on file.

In order to enhance the tolerance to packet loss and bit errors, the sender deploys the derived Walsh codes to modulate the digital watermarks. Some synchronization marks, denoted as SYNC, are inserted into the modulated Walsh codes sequence. That is 1111 stands for SYNC, 1001 stands for ‘0’, and 1010 stands for ‘1’. The resulting sequence is represented as $\{b_0, b_1, b_2, b_3, b_4, b_5, b_6, b_7,$ **SYNC**, ….., SYNC, $b_{j-1}, b_j, b_{j+1}, …. b_n\}$. The bits in between an adjacent SYNC pair consist of a phase. The sender embeds this bit sequence into the outgoing multimedia stream.

On the other side, the receiver retrieves a bit sequence, $\{b_0', b_1', b_2', b_3'$, SYNC, ….., SYNC, $b_{j-1}', b_{j+1'}, …. b_n'\}$, from the incoming stream. Since both the sender and the receiver share *comm_watermark*, the reference bit sequence, $\{b_0, b_1, b_2, b_3$, SYNC, ….., SYNC, $b_{j-1}, b_j, b_{j+1}, …. b_n\}$, is accessible by the receiver.

During the communication procedure the receiver performs the following steps to evaluate the integrity and authenticity of the data being transmitted in the channel:

```
Initialisation: prepare a buffer to hold
the retrieved bits
Step:
    1) scan buffer; if SYNC found, then
       goto step 1
    2) extract one embedded digital
       watermark bit from one multimedia
       frame and put it into buffer; repeat
       step 1
    3) evaluate the partial integrity and
       authenticity; clear the buffer
    4) if not end of communication, then
       goto step1.
```

Due to the possible packet loss and bit errors, the integrity and authenticity assessment looks a little complicated.

(1) If the size of the extracted bit sequence (N) is equal to the size of the reference bit sequence (no packet loss has happened), then compare the extracted bit sequence with the reference bit sequence as follow:

$$I = \frac{\sum_{i=1}^{N} (b_i \text{ XNOR } b_i')}{N} \qquad \textbf{(2)}$$

(2) Maybe the size of the extracted bit sequence, M, is not equal to the size of the reference bit sequence, N, either due to packet loss or due to a loss of SYNC. If the size of the extracted bit sequence is smaller, then the integrity and authenticity estimation work as follows:

```
Denote the reference bit sequence as $B_r$ and
the extracted bit sequence as $B_x$
Step:
    1) find the most left and the most
       right common strings (sizes are de-
       noted as $l_l$ and $l_r$) from $B_r$ and $B_x$
    2) remove the most left and the most
       right common strings from the two
       bit sequences
    3) find the longest common string from
       the two bit sequences; the length of
       the common string is $l_c$
```

The integrity value of this phase is given by

$$I = \frac{l_l + l_r + l_c}{N} \qquad \textbf{(3)}$$

(3) If the size of the extracted bit sequence is larger (the synchronization code has been destroyed), then combine the adjacent two reference phases and perform the above procedure to estimate the multimedia integrity and authenticity.

Many existing real-time multimedia communication protocols deploy symmetric ciphers to set up a secure channel. In these cases the proposed real-time integrity and authenticity assessment can be exploited as a method to measure the quality of the transmission.

2.4 Source Origin Authentication

The secret digital watermarks, *speech_-watermarks*, embedded in the multimedia messages have been signed by *transServer* using public-key encryption. This means that each of the watermarked conversation data is distinguishable from others. Therefore, one can verify the source origin of the saved multimedia data: 1) A (*speech_watermark*, *key*) pair can be parsed from the proper *transCert*; 2) extract a digital watermark, *w*', from the recorded conversation data and compare it with the original one to authenticate the inspected speech.

3. Speech Watermarking Algorithm Incorporated with GSM 610 Coders

3.1 GSM 610 Voice Encoder

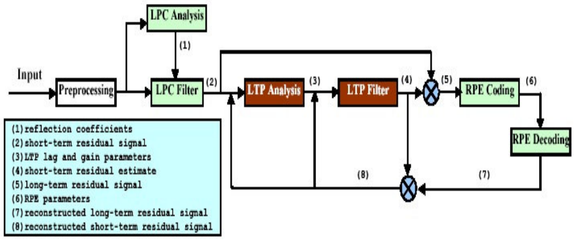
LPC Analysis
(1)
Input
Preprocessing
LPC Filter
(2)
LTP Analysis
(3)
LTP Filter
(4)
(5)
RPE Coding
(6)
RPE Decoding
(8)
(7)
(1)reflection coefficients
(2)short-term residual signal
(3)LTP lag and gain parameters
(4)short-term residual estimate
(5)long-term residual signal
(6)RPE parameters
(7)reconstructed long-term residual signal
(8)reconstructed short-term residual signal

Fig. 7 GSM 610 Voice Encoder

The GSM coder [13] as in Fig. 7 is being widely used in current wireless telecommunication systems. The vocoder model consists of a tone generator (which models the vocal chords) and a filter that modifies the tone (which in turn models the mouth and nasal cavity shape). It first removes the local correlation between samples and produces either a pitch residual or a noise-like signal for unvoiced speech, and then quantifies the harmonics of the speech. Later, the RPE stage reduces the residual samples down to four sets of 13 bit sub-sequences. The optimum sub-sequence is determined as having the least error value. The resulting signal is fed back in order to help the processing of the next frame. The resulting signal structure is shown in Table 1.

Table 1. Output of GSM 610 Voice Encoder

Parameter	Number of values	Bits per frame
LARs	8 per frame	36 bits
LTP lag	1 per subframe (7 bits)	28 bits
LTP gain	1 per subframe (2 bits)	8 bits
RPE grid position	1 per subframe (2 bits)	8 bits
Block amplitude	1 per subframe (6 bits)	24 bits
RPE Pulses	13 per subframe (3 bits each)	156 bits
Total		**260 bits**

3.2 A Fragile Watermark Algorithm Incorporating the GSM 610 Voice Encoder

In the proposed scheme RPE pulses are modified slightly to engrave the digital watermarks as shown in Fig. 8. As already mentioned, two digital watermarks, *speech_watermark* and *comm_-watermark*, are used for each transaction. Both the marks and the private keys are binary sequences.

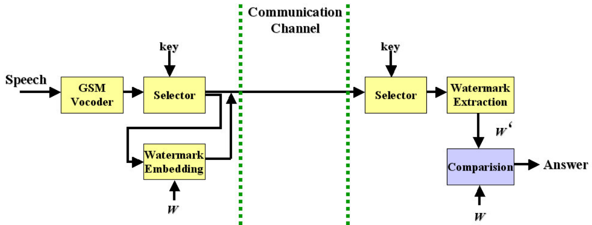

Fig. 8 Speech Watermarking Scheme

A GSM 610 coder generates 52 RPE pulses of 3 bits each for a speech frame of 20 ms duration. An RPE pulse pair in one frame is chosen to embed one bit of the digital watermark. The pulse pair is selected according to *key* as depicted in Fig. 8.

Consider x’ and x to be the least significant bits, respectively, of the selected candidate RPE pulses. Then the watermark embedding and extraction algorithm executes as follows:

```
Embedding                              Extraction
E(x', x, w)                            EX(x', x)
{                                      {
  if (|x'- x| ≠ w )  x←¬x;               return |x'-x| ;
}                                      }
```

(1) *speech_watermark*

key is an integer pair {(p1, p2) | 0≤ p1, p2<50}. To embed one bit of the secret watermark, *speech_watermark*, select two RPE pulses from the first 50 RPE pulses according to *key*. The least significant bit of two selected RPE pulses is the candidate bit pair (x, x’). Apply the algorithm mentioned above to engrave a bit of *speech_watermark.*

(2) *comm_watermark*

comm_watermark is public.The last two RPE pulses of one frame are selected to hide one bit of *comm_watermark*. The least significant bits of each selected RPE pulses are the candidate bit pair (x, x’). Apply the algorithm mentioned above accordingly. Fig. 9 shows an example of how the proposed audio watermarking algorithm operates.

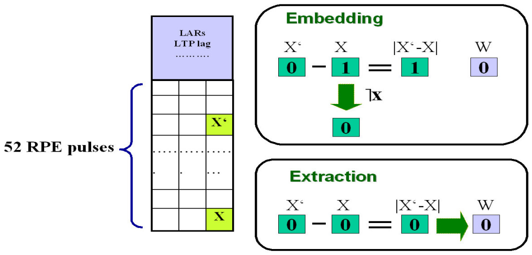

Fig. 9 Audio Watermarking Example

3.3 Speech Feature Extraction

Previous research for content integrity verification resulted in the feature extraction method. The feature extraction method has similar functionalities for multimedia data [8, 9, 10, 11, 12] as the hash function for text messages. The technique of self-embedding is a kind of feature extraction skill, which tries to embed extracted feature values back into the audiovisual data by means of high data capacity watermarking. Wu’s work [5] shows that self-embedding has two drawbacks: (a) the watermarking algorithm needs a very high data capacity, (b) most self-embedding methods operate at high computational costs. Thus the known self-embedding algorithms are not applicable for low data rate speech encoders such as GSM 610 voice encoder. Therefore, we developed a new feature-embedding algorithm for such encoders.

Each person is uniquely identified by her/his specific speech pattern. Log Area Ratios (LARs) are the reflection coefficients modeling the shape of a vocal tract, such that it is not surprising to exploit this technique in speaker recognition [16]. So, we selected LARs as the principal speech features too. A GSM 610 coder generates 36 bits LAR parameters for every 20 ms speech duration. Its feature value can be estimated as

$$F(\, LARs \,) = \mathrm{MOD}\left(\sum_{i=1}^{36} \mathrm{LAR}_i \,/\, 2 \right). \qquad \textbf{(4)}$$

We then apply the function $\Theta(\, w,\, F(\cdot)\,)$ to generate the new watermark bit *w’*. Hence, the speech feature is integrated into the proposed speech watermarking algorithm.

Table 2. Function Θ(*w*, *F(·)*)

		F(LARs)	
		0	**1**
W	**0**	**1**	**0**
	1	**0**	**1**

The watermark detection can apply the outlined procedure to verify the integrity of the speech. Both transmission errors and attacks can degrade the integrity value, *I*, of the speech data.

4. Experimental Results

We performed some test runs for the proposed speech watermarking algorithm. It is obvious from Table 3 that the trade-off of the proposed speech watermarking approach is rather low.

Table 3. Watermarking Overhead on GSM 610 Voice Encoder

Encoding/Decoding Time of a Frame	Computing Time for Watermarking in a Frame	Watermarking Time/Encoding Time
287.63 µs	2.79 µs	1 %

Pentium 3 - 800M Hz CPU and 356 MB RAM

The proposed speech watermarking algorithm slightly modifies some niches of RPE pulses values, hence it does not degrade the overall quality of speech. As an example, Fig. 10 visualises some original and watermarked speech samples.

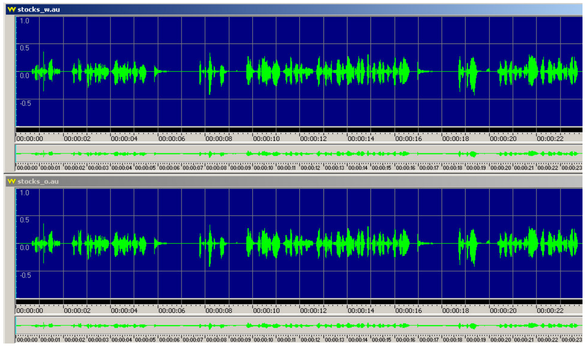

Fig. 10 Watermarked vs. Original Speech Samples

To prove the evidence of the proposed scheme, a prototype has been developed and implemented by means of a desktop computer and a notebook . The prototype architecture consists of a pair of IP phones and a *transServer*. IP phones are adapted from Speakfreely [14], a well-known open source IP telephony implementation. Both the implemented *transServer* and IP phones employ the OpenSSL toolkit [15] to handle digital certificates and to generate *transCert*s. The first experiment was performed on a 100M Ethernet LAN being used as the communication network.

In addition, we tested the network scenario comprising both an 100Mb/s Ethernet LAN and a 11Mb/s WLAN. The upper row of Table 4 shows the result of a test, which was run by a VoIP end-point couple on 100Mb/s Ethernet. The lower row shows the test result taken from the wireless LAN scenario as described in Section 1. The average packet loss (including packet delay) is quite low for VoIP devices working on wired LAN and the performance decreases in a wireless environment. But the proposed integrity measurement is robust enough to suppress the degradation of the network.

Table 4. Integrity Values

	No. Total Packet	**No. Packet Delay & Errors**	**No. Water -mark Errors**	**Integrity Value**
Test on 100Mb/s Ethernet LAN	16390	6	6	99.96%
Test on 11Mb/s WLAN	16908	285	112	99.83%

$$I_{\mathrm{wLan}} = \frac{[((16908/4)–112)*1] + [112*0.75]}{16908 / 4}$$

$$= 99.83\%$$

References:

[1] IEEE Draft International Standards, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, *IEEE P801.11/D10*, 1999.

[2]http://www.tmcnet.com/articles/itmag/0199/0199roundt.htm.

[3] M. Baugher et al., The Secure Real Time Transport Protocol, *IETF Draft*, 2002.

[4] Z. Li, R. Xu, Energy Impact of Secure Computation on a Handheld Device, *IEEE International Workshop on Workload Characterization*, pp.109 –117, 2002.

[5] C. Wu and C. Kuo. Speech Content Integrity Verification Integrated with ITU g.723.1 Speech Coding, *IEEE International Conference on Information Technology: Coding and Computing*, pp. 680–684, 2001.

[6] J. Haitsma, M. Veen, T. Kalker and F. Bruekers, Audio Watermarking for Monitoring and Copy Protection, *Proceedings of the 2000 ACM workshops on Multimedia*, pp.119–122, 2000.

[7] D. Gruhl, W. Bender, and A. Lu, Echo-hiding, *Information Hiding: 1st International Workshop*, pp.295–315, 1996.

[8] D. Lou and J. Liu, Fault Resilient and Compression Tolerant Digital Signature for Image Authentication, *IEEE Transactions on Consumer Electronics*, pp.31–39, 2000.

[9] C. Rey and J. Dugelay, Blind Detection of Malicious Alterations on Still Images Using Robust Watermarks, *IEE Seminar on Secure Images and Image Authentication*, pp.7/1 –7/6, 2000.

[10] M. Schneider and S. F. Chang, A Robust Content Based Digital Signature for Image Authentication, *Proceedings of IEEE International Conference on Image Processing (ICIP'96)*, pp. 227–230, 1996.

[11] M. Wu and B. Liu, Watermarking for Image Authentication, *Proceedings of International Conference on Image Processing*, pp.437–441, 1998.

[12] J. Dittmann, A. Steinmetz, and R. Steinmetz, Content-based Digital Signature for Motion Pictures Authentication and Content-fragile Watermarking, *IEEE International Conference on Multimedia Computing and Systems*, pp. 209–213, 1999.

[13] GSM Vocoder Specification, *ITU*, 1996.

[14] Brian C. Wiles & John Walker, *Speakfreely*.

[15] http://www.openssl.org.

[16] J. Campbell, Speaker Recognition, *Proceedings of IEEE*, Vol. 85, No.9, pp.1437-1462, 1997.