Watermark-Based Authentication and Key Exchange in Teleconferencing Systems

Ulrich Rührmair[a], Stefan Katzenbeisser[b],
Martin Steinebach[c], and Sascha Zmudzinski[c]

[a]Technische Universität München, Department of Computer Science
[b]Technische Universität Darmstadt, Computer Science Department
[c]Fraunhofer Institute for Secure Information Technoology SIT, Darmstadt

Abstract. In this paper we propose an architecture which combines watermarking with traditional cryptographic key agreement protocols to establish an authenticated or encrypted channel in teleconferencing systems. Technically the proposed method embeds messages of the key agreement protocol within an audio or video stream and is based on the assumption that the human communication partners can recognize each other easily; the watermark establishes a close coupling between the cryptographic key exchange messages and the media stream. We argue that the security of the scheme is based on a yet unexplored security property of digital watermarks; furthermore we present preliminary research results that suggest that this property holds in standard watermarking schemes.

1 Introduction

After the introduction of public-key cryptography, the problem of secure communication between two parties who have never exchanged secret keys seemed to be solved in practice. Ideally, communication partners access a trusted register (similar to a telephone book) in order to download each others public keys, which should be maintained by a trusted certification authority. Subsequently, the parties can use this key in order to encrypt and authenticate messages. Over three decades later, it has turned out that realizing this vision through a worldwide Public Key Infrastructure (PKI) is by no means simple. In fact, all attempts to establish a large-scale PKIs have failed due to organizational problems and lack of trust in certification authorities [2, 3].

It is thus pressing to consider alternative ways of authentication and key exchange, which do not utilize a PKI, but still allow to establish an encrypted and authenticated channel between two parties. In this paper, we consider a teleconferencing scenario, where two parties use digital telephony or a video conferencing system to establish a connection. Often, both communication partners know each other well enough to recognize their voices and/or each other on the transmitted audio or video stream. We propose to utilize this human knowledge and mutual recognition ability to authenticate a channel; key transfer is realized by using a watermark, which is directly embedded in this channel. More precisely, we embed messages of a key agreement protocol inseparably into the transmitted audio or video signal by means of a robust watermark. The analog voice or video signals, which can mutually be recognized as genuine by the communication partners, provide an authenticated channel that inhibits real-time masquerading attacks.

A natural way of establishing an encrypted phone or video conference utilizing the approach depicted above would work as follows: In the first phase, conversation is only transmitted in plain, but carries embedded information allowing secure key exchange. The messages required for key exchange are embedded by a watermark in the content. The two communication partners identify each other by recognizing the transmitted audio or video stream, and thereby implicitly authenticate the embedded key exchange messages. Once the key exchange has been completed, the system can encrypt all further communication between Alice and Bob in a second phase. The method seems suitable for establishing encrypted telephone or video communication in a highly convenient, ad hoc fashion, and would readily be usable with existing technology such as *skype* or similar services.

The approach is in principle similar to the Cryptophone architecture [1], where the communication equipment used by both communication partners sends Diffie-Hellman key exchange messages to each other. To detect a potential man-in-the-middle attack, the phones generate digests of the sent messages, which are displayed to the users. Both communication partners read the digest displayed on the phone over the encrypted line and the partner verifies if the hash was read correctly. If any discrepancy is detected, potentially a man-in-the-middle attack occurred. In comparison to the existing Cryptophone architecture, our approach is much more convenient as it does not require reading hash values aloud but ensures authenticity in the background of the communication.

The rest of the paper is organized as follows. In Section 2 we propose a protocol that realizes the above mentioned authentication method in audio or video signals. Subsequently, Section 3 discusses the security of the proposed protocol; in particular we analyze which security property the underlying watermarking scheme needs to have to make the overall scheme secure.

2 Authentication Protocol

In this section we propose a protocol for teleconferencing systems that authenticates messages of key agreement protocols by help of the human user. As an example we describe the protocol in conjunction with the Diffie-Hellman key exchange.

The protocol utilizes a watermarking scheme which needs to be robust against varying channel conditions (the recipient will abort the protocol if no watermark is detectable in the signal). Instead, we use a different security assumption, which is detailed in the next section.

1. Both parties share the public system parameters: a random prime p, a generator g of $\mathbb{Z}_p^*$ and a global watermarking key K_W. These parameters can e.g. be distributed along with the communication software.
2. Alice chooses suitably and uniformly at random an exponent a.
3. Alice and Bob begin their telephone or video call.
4. Alice embeds in her part of the conversation during the first seconds a watermark by use of the watermarking key K_W. This watermark contains the payload $H(p,\, g,\, g^a \ mod\ p)$, where H denotes any cryptographic hash function. Furthermore, Alice sends the tuple $(p,\, g,\, g^a \ mod\ p)$ alongside the communication.

5. Bob verifies whether Alice's transmission was watermarked with key K_W, that the watermark payload is a hash of the sent key exchange message, and that he is speaking to Alice by recognizing her image/voice. Furthermore, Bob also uses a detector software to scan for artifacts of signal manipulations, similar to tools used in forensic investigations (see Section 3). If any test fails, or if he did not receive a watermarked content, he aborts the protocol.
6. Bob chooses an exponent b and embeds a watermark with payload $H(g^b)$ in the data stream (simultaneously to Alice). Furthermore, Bob sends the information g^b to Alice.
7. Analogously to Bob, Alice verifies whether Bob's transmission was watermarked with key K_W, that the watermark payload is the hash of the key exchange message and that she is speaking to Bob and by recognizing his image/voice. She performs the same tests as Bob to detect any tampering with the audio or video stream. If any test fails, or if she did not receive watermarked content, she aborts the protocol.
8. Alice and Bob set their joint secret key to be $K = g^{ab}$.
9. Alice and Bob henceforth encrypt their communication with the key K. Optionally, Alice and Bob may send each other confirmation messages that are authenticated with the key K to verify that the key exchange was successful.

The described mechanism for authenticating messages via watermarks and voice and/or image signals does not necessarily need to be used in connection with the Diffie-Hellman key exchange protocol. Any other key exchange protocol that operates in one challenge-response round is also possible; in this case, the public Diffie-Hellman messages are replaced with the messages of the key exchange protocol.

Instead of authenticating the messages of a key exchange protocol, one can also directly exchange and authenticate the public keys of the communication partners via the described mechanism in the audio and/or video streams. As usual, these transferred public keys can then subsequently be used for encryption and authentication of further speech or video signals.

As a further variant, one could imagine that communication partners record little voice and/or video samples, embed their public key in the sample via a watermarking scheme, and post these sequences on their web-pages. Others can download these samples, watch them to judge whether they are genuine, and recover the embedded public key. These sequences could also be recorded newly in short distances, for example on a daily basis, with recorded person pronouncing the current date. This would not overstretch the security properties of the watermarks: They merely had to remain secure for one day in that example.

3 Security

Since the underlying key exchange mechanism is known to be secure, attacks mainly stem from the watermarking layer. It is crucial to note that, due to the fixed symmetric watermarking key K_W, we have to assume that this key is available to an adversary as well. In principle we can distinguish between passive and active attackers; the former record the exchanged communication and try to obtain the key K, while the latter replace messages in order to impersonate one communication partner.

Passive attacks. The system inherits its security against passive attacks from the use of the Diffie-Hellman protocol. That is, an eavesdropper will not be able to restore K, unless he can break the cryptographic part of the protocol. However, since Diffie-Hellman key exchange provides no security against active attackers, security against active attackers must be considered separately.

Active attacks: Replaying unmarked sequences. In this type of attack, an adversary tries to obtain unmarked video or audio sequences of the communication partner (such as fragments of speech posted on web sites or captured by analog recording devices), selects its own Diffie-Hellman message and embeds the message into the signal by using its knowledge of the watermarking key K_W. In an even more sophisticated system, an adversary may even assemble speech fragments of one communication partner in order to generate a new coherent speech signal. Note that the same attack is possible against the Cryptophone architecture (record the voice of the communication partner when he pronounces the individual characters of the message digest and use speech synthesis to create a speech signal for a new digest). This attack cannot be prevented by technical means, but must be tackled by the communication partners themselves. For example, both communication partners can try to individualize the initial (unencrypted) part of the communication, e.g. by saying the time, the date, or by having Alice and Bob pronounce randomly chosen, unusual words from a dictionary. Alternatively, Alice and Bob might ask each other questions that only the correct communication partner can answer, or take similar measures to assure that the communication is not replayed or modified. Note that prevention of this attack in our scenario seems to be easier than in the Cryptophone architecture, since it is considerably easier to automatically synthesize speech pronouncing a small set of message digests rather than complex text.

Active attacks: Masquerading Attacks. Another threat consists in an adversary who records one protocol run and replays it at a later time with a different watermark. If successful, such an attack would allow to alter the watermark payload (to point to the key of the attacker), while the audio file is still authenticated by the recipient. This attack is particularly problematic, since many phone calls resemble each other in the first part (i.e., the parties state their names, greet each other, etc.). To exploit this, Eve records one typical initial conversation of Alice and re-embeds her own watermark into the data stream (potentially after performing signal processing operations aiming at removing the first watermark). Subsequently she can mount a masquerading attack by replaying the newly watermarked sequence.

Protection against this attack requires a novel security property of the employed digital watermark: a communication partner must be able to decide whether the watermark was embedded in a previously unmarked media object or in one that already carried a watermark that was embedded with the same key; first results can be found in [5]. For the security of the scheme, both parties thus need to perform an analysis of the audio or video stream to detect such manipulation attempts. In contrast to the re-marking problem (where several watermarks are embedded with different keys), the particular problem has not been discussed in the literature yet.

By applying methods from media forensics, it is possible to distinguish (within some error bounds) whether a media file has been marked before with the same key,

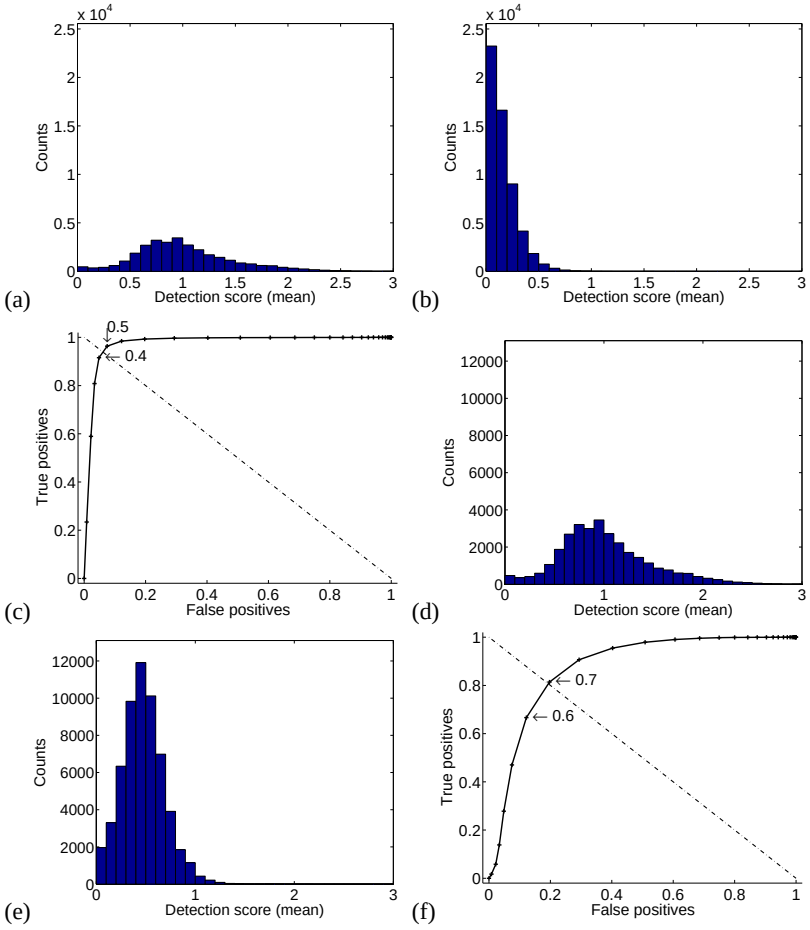

Fig. 1. Robustness against re-marking with the same key.

since every watermarking process irreversibly changes the statistical characteristics of a media object. To see whether this is possible using a standard watermarking scheme we performed some initial experiments. In particular, we used an audio watermarking system, which employs the Patchwork embedding method in the Fourier domain and is used in commercial products [4, 5]. We tested different (mono) audio material of approximately 8.5 hours length, sampled at a rate of 44.1 kHz, into which we embedded a watermark sequence; subsequently we attacked the content in order to make a watermark visible where all payload bits are flipped. If a second watermark is embedded at the same strength as the first one (3dB in our experiment), the histograms of the detection statistics within single (Figure 1(a)) and doubly marked content (Figure 1(b)) are clearly distinguishable by a simple classifier which compares the value of the detection response by a threshold. The ROC curve of this classifier is shown in Figure 1(c); this plot shows that both cases can be distinguished with an EER of approximately 7%. If

an attacker embeds the second watermark with a larger strength (6dB), the two cases can still be distinguished (Figures 1(d) and 1(e) show the histograms of single and doubly marked content, while Figures 1(f) depicts the ROC curve), albeit at larger error rates (EER of approximately 20%). In addition, closer analysis shows that the adversary has to accept significant losses in sound quality, which might signal the attack to the communication partner. These preliminary results indicate that by using appropriate forensic analysis of the watermark detection responses, it is in principle possible to automatically determine whether a signal underwent malicious re-marking attacks under the same key. Further research is ongoing to determine the susceptibility of this classifier to other signal processing attacks.

4 Conclusions

We described a protocol that allows key establishment in digital telephony or video conferencing applications. In particular, we embed messages of a key exchange protocol as watermark in the audio or video stream and rely on the communication partners to recognize their voices. Due to the coupling of the messages and the digital data stream, the source of the messages is authenticated. We showed that the security of the approach crucially depends on a novel security property of watermarks, which has not yet been discussed in the literature: one should be able to distinguish an object where a watermark was embedded in a previously unmarked media from one that already carried a watermark. Initial experiments showed that this question can in principle be answered (using a standard watermarking scheme) by analyzing the statistical properties of the detection response. Future work thus includes the definition of forensic methods that allow to answer this question with lower error rates.

References

1. `http://www.cryptophone.de`
2. Carl Ellison, Bruce Schneier, Ten Risks of PKI: What You're not Being Told about Public Key Infrastructure, in *Computer Security Journal*, vol. XVI, no. 1, 2000.
3. Carl Ellison, Bruce Schneier: Risks of PKI: Secure Email, in *Communications of the ACM*, vol. 43, no. 1, p. 160, 2000.
4. M. Steinebach, S. Zmudzinski, Evaluation of robustness and transparency of multiple audio watermark embedding, in *Proceedings of the SPIE, Security, Steganography, and Watermarking of Multimedia Contents X*, 2008
5. M. Steinebach, S. Zmudzinski, S. Katzenbeisser, U. Rührmair, Audio watermarking forensics: detecting malicious re-embedding, to appear in *Proceedings of SPIE Vol. 7541, Media Forensics and Security XII*, 2010.