MULTIMEDIA CONTENT AUTHENTICATION: FUNDAMENTAL LIMITS

Emin Martinian and Gregory W. Wornell

Dept. Electrical Engineering and Computer Science
Massachusetts Institute of Technology, Cambridge, MA 02139
{emin,gww}@mit.edu

ABSTRACT

In many multimedia applications, there is a need to authenticate a source that has been subjected to benign degradations in addition to potential tampering attacks. We develop a meaningful formulation of this problem, and identify and interpret the associated information-theoretic performance limits. The associated systems are shown to perform dramatically better than frequently proposed approaches based on information embedding techniques.

1. INTRODUCTION

In traditional authentication problems, the goal is to determine whether a received message is an exact replica of what was sent. Digital signature techniques are a natural tool for addressing such problems. However, in many emerging multimedia applications, the message may be an audio or video waveform, and even in the absence of a tampering attack, the waveform may experience routine degradation due to noise, compression, etc., before being received. Methods for reliably authenticating the received data in such cases are important as well—see, e.g., the references in [1].

As a motivating example, consider the authentication of drivers' licenses. Many jurisdictions print a hologram on the photograph portion of the license. The presence of the hologram indicates that the license is legitimate without adding excessive visual distortion. Imprinted holograms are a particular implementation of a larger class of schemes that use special markings embedded into the photograph to enable a decoder to extract an authentic representation of the original. In general, the special markings should be embedded so that the distortion between the original and embedded photographs is small, thus enabling a receiver without an appropriate decoder to still use the license to check the identity of the bearer. In addition, the special markings need to be robust to perturbations in the form of smudges, scratches, or other degradation due to routine handling: the decoder should still be able to authenticate if only these are present. Finally the special markings should be inserted so that no other agent can create a successful forgery.

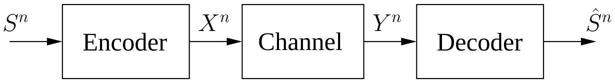

Fig. 1. Authentication system model. The source S^n is encoded to create the channel input X^n, incurring some distortion. The channel models benign degradations due to routine handling, as well as tampering by a malicious attacker. The decoder produces from the channel output Y^n either an authentic reconstruction $\hat{S}^n$ of the source to within some fidelity, or indicates that authentication is not possible.

This paper examines key aspects of the fundamental tradeoffs between security, robustness, and distortion using information-theoretic analysis. Section 2 describes the model for authentication. Section 3 presents a formal problem statement and characterizes the achievable performance of authentication systems. Section 4 applies the results to the practically important Gaussian-quadratic scenario and quantifies the improvement over traditional information embedding approaches, and Section 5 contains some concluding remarks.

2. SYSTEM MODEL AND PROBLEM FORMULATION

Our system model is as depicted in Fig. 1. To simplify the exposition, we model the original source as an independent and identically distributed (i.i.d.) sequence $S_1, S_2, \ldots, S_n$ denoted as S^n. In practice S^n could correspond to sample values or signal representations in some suitable basis.

The encoder takes as input a block of n source samples S^n, producing an output X^n that is suitably close to S^n with respect to some distortion measure. The encoded signal then passes through a channel, which captures the effects of routine handling as well as any tampering, producing the channel output Y^n.

The decoder either produces, to within some fidelity as

This work has been supported in part by NSF under Grant No. CCR-0073520, Hewlett-Packard through the MIT/HP Alliance, and Texas Instruments through the Leadership Universities Program.

To appear in *Proc. ICIP-2002*, September 22-25, 2002, Rochester, NY.

quantified by a suitable distortion measure, a reconstruction $\hat{S}^n$ of the source that is guaranteed to be free from the effects of any tampering by an attacker, or declares that it is not possible to produce such a reconstruction. We term such reconstructions "authentic."

In the authentication scenario of Fig. 1, we deliberately avoid a particular channel model. Rather, we make use of a concept we refer to as a "reference channel." A reference channel, or more accurately a reference channel ensemble, is the collection of realized channels for which we require the decoder to produce an authentic reconstruction of the source. For channels outside this collection, we allow the decoder to declare that an authentic reconstruction is not possible. Thus, the reference channel should be chosen to capture all benign and malicious degradations we desire the system to overcome.

Given a particular reference channel (known to the transmitter, receiver, and attacker), the goal of the system designer is to make the encoding distortion small, so that in the absence of a channel the encoder output is a faithful replica of the original source, and to make any authentic reconstructions the decoder produces of high fidelity. In general, these are conflicting objectives, and in the sequel we explore the fundamental trade-offs involved.

The following definition makes precise the notion of an authentic reconstruction, i.e., one free from the effects of the channel.

Definition 1 *A reconstruction $\hat{S}^n$ produced by the decoder from the output Y^n of the channel is said to be authentic if it satisfies the following Markov condition:*

$$\hat{S}^n \leftrightarrow \{S^n, X^n\} \leftrightarrow Y^n \tag{1}$$

This condition is rather intuitive: it means that the channel output cannot influence an authentic estimate directly, but rather only indirectly through the encoder inputs and outputs.

As our main result, in Section 3 we characterize when authentication systems are possible, and when they are not. Specifically, let D_e denote the encoding distortion, i.e., the distortion experienced in the absence of a channel, and let D_r denote the distortion in the reconstruction produced by the decoder when the signal can be authenticated, i.e., when the channel degradations are consistent with the reference distribution $p(y|x)$. Then we determine which distortion pairs $(D_\mathrm{e}, D_\mathrm{r})$ are asymptotically achievable.

Before developing our main result, we illustrate with an example the character of results that are obtained.

Preliminary Example An illustrative achievable distortion region is depicted in Fig. 2. This example, developed in [1], corresponds to a problem involving a symmetric Bernoulli source, Hamming distortion measures, and a binary symmetric reference channel with crossover probability p. When authentication is not required, all points above the dashed line can be achieved. Thus requiring authentication strictly reduces the set of achievable distortions when $p > 0$. Note that at the point $(D_\mathrm{e}, D_\mathrm{r}) = (p, p)$, the decoder completely eliminates the effects of the reference channel when it is in effect: the minimum achievable reconstruction distortion D_r is the same as the distortion D_e at the output of the encoder. Observe, too, that the case $p = 0$ corresponds to the traditional scenario for digital signatures where there is no noise. In this case, as the figure reflects, authentication is achievable without incurring any encoding distortion nor reconstruction distortion.

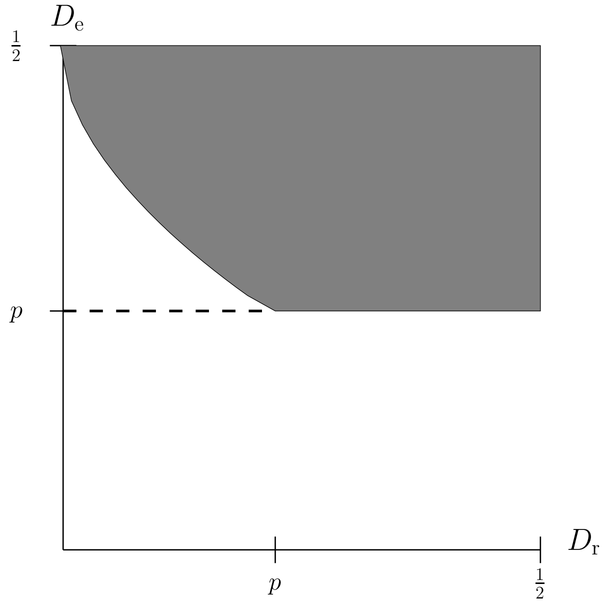

Fig. 2. The achievable distortion region for a symmetric Bernoulli source transmitted over a binary symmetric reference channel with crossover probability p. Distortions are with respect to the Hamming measure. The case $p = 0$ corresponds to traditional digital signatures.

3. CHARACTERIZATION OF SOLUTION: CODING THEOREMS

An instance of the authentication problem consists of the tuple

$$\{\mathcal{S}, p(s), \mathcal{X}, \mathcal{Y}, p(y|x), d_\mathrm{e}(\cdot,\cdot), d_\mathrm{r}(\cdot,\cdot)\} \tag{2}$$

where $\mathcal{S}$, $\mathcal{X}$, and $\mathcal{Y}$ are the source, channel input, and channel output alphabets—which are finite unless otherwise indicated—and $p(s)$ is the (i.i.d.) source distribution, $p(y|x)$ is the (memoryless) reference channel law, and $d_\mathrm{e}(\cdot,\cdot)$ and $d_\mathrm{r}(\cdot,\cdot)$ are the encoding and reconstruction distortion measures, respectively.

A solution to this problem (i.e., an authentication scheme) consists of an algorithm that returns an encoding function

Υ^n, a decoding function Φ^n, and a secret key.[1] The secret key is shared only between the encoder and decoder; all other information is known to all parties including attackers. The encoder is a mapping from the source sequence (as well as the secret key, which we suppress to simplify notation) to codewords, i.e., $\Upsilon^n(S^n) : \mathcal{S}^n \mapsto \mathcal{X}^n$.

The decoder is a mapping from the channel output and the secret key to either an authentic source reconstruction $\hat{S}^n$ (i.e., one satisfying (1)) or the special symbol $\varnothing$ that indicates such a reconstruction is not possible: $\Phi^n(Y^n) : \mathcal{Y}^n \mapsto \mathcal{S}^n \cup \{\varnothing\}$. Notice that since an authentic reconstruction must satisfy (1), and since the decoder must satisfy the Markov condition $\{S^n, X^n\} \leftrightarrow Y^n \leftrightarrow \Phi^n(Y^n)$, we have that $\hat{S}^n \leftrightarrow \{S^n, X^n\} \leftrightarrow \Phi^n(Y^n)$ forms a Markov chain. Therefore the authentic reconstruction is effectively defined by the encoder and not the decoder.

The relevant distortions are the encoding and decoding distortion computed as the sum of the respective (bounded) single letter distortion functions d_e and d_r, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} d_\mathrm{e}(S_i, X_i) \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} d_\mathrm{r}(S_i, \Phi_i^n(Y^n)).$$

The system can fail in one of three ways. The first two failure modes correspond to either the encoder introducing excessive encoding distortion, or the decoder failing to produce an authentic reconstruction with acceptable distortion when the reference channel is in effect. For any $\epsilon > 0$, the associated distortion violation error events are

$$\mathcal{E}_{D_\mathrm{e}} = \left\{ \frac{1}{n}\sum_{i=1}^{n} d_\mathrm{e}(S_i, X_i) > D_\mathrm{e} + \epsilon \right\} \tag{3}$$

$$\mathcal{E}_{D_\mathrm{r}} = \left\{ \Phi^n(Y^n) = \varnothing \right\} \cup \left(\left\{ \Phi^n(Y^n) \neq \varnothing \right\} \cap \left\{ \frac{1}{n}\sum_{i=1}^{n} d_\mathrm{r}(S_i, \Phi_i^n(Y^n)) > D_\mathrm{r} + \epsilon \right\} \right). \tag{4}$$

In the remaining failure mode, the system fails to produce the desired authentic reconstruction $\hat{S}^n$ from the channel output and, instead of declaring that authentication is not possible, produces an incorrect estimate. The successful attack event is

$$\mathcal{E}_\mathrm{sa} = \{\Phi^n(Y^n) \neq \varnothing\} \cap \{\Phi^n(Y^n) \neq \hat{S}^n\}. \tag{5}$$

Definition 2 *The achievable distortion region for the problem* (2) *is the closure of the set of pairs* $(D_\mathrm{e}, D_\mathrm{r})$ *such that there exists a sequence of authentication systems, indexed by* n*, where for every* $\epsilon > 0$ *and as* $n \to \infty$, $\Pr[\mathcal{E}_\mathrm{sa}] \to 0$ *regardless of the channel law in effect,* $\Pr[\mathcal{E}_{D_\mathrm{e}}] \to 0$*, and* $\Pr[\mathcal{E}_{D_\mathrm{r}}] \to 0$ *when the reference channel is in effect, with* $\mathcal{E}_\mathrm{sa}$, $\mathcal{E}_{D_\mathrm{e}}$ *and* $\mathcal{E}_{D_\mathrm{r}}$ *as defined in* (5), (3), *and* (4).

For such systems, we have the following theorem [1]:

Theorem 1 *The distortion pair* $(D_\mathrm{e}, D_\mathrm{r})$ *lies in the achievable distortion region for the problem* (2) *if and only if there exists a distribution* $p(u|s)$ *and functions* $f(\cdot,\cdot)$ *and* $g(\cdot)$ *such that*

$$I(U;Y) - I(S;U) \geq 0 \tag{6a}$$

$$E[d_\mathrm{e}(S, f(U,S))] \leq D_\mathrm{e} \tag{6b}$$

$$E[d_\mathrm{r}(S, g(U))] \leq D_\mathrm{r} \tag{6c}$$

where the alphabet $\mathcal{U}$ *of the auxiliary random variable* U *generated by* $p(u|s)$ *has cardinality*[2] *bounded by* $|\mathcal{U}| \leq (|\mathcal{S}| + |\mathcal{X}| + 3) \cdot |\mathcal{S}| \cdot |\mathcal{X}|$.

3.1. Layered Authentication Systems

It is also possible [1] to develop layered systems where the decoder produces a coarse-grain reconstruction $\hat{S}^n$ when the received signal is consistent with the channel law $p(y|x)$, or a fine-grain reconstruction $\hat{S}_\mathrm{f}^n$ when the received signal is consistent with the reference channel law $p(z|x)$. For (degraded broadcast) reference channel laws of the form $p(y,z|x) = p(y|z)p(z|x)$, the distortion triple $(D_\mathrm{e}, D_\mathrm{r}, D_\mathrm{r}^\mathrm{f})$ is achievable if there exists a conditional distribution $p(u,t|s)$, and scalar decoding functions $g(\cdot)$ and $g(\cdot,\cdot)$ such that

$$I(T;Y_\mathrm{f}|U) - I(S;T|U) \geq 0 \tag{7a}$$

$$E[d_\mathrm{r}^\mathrm{f}(S, g(U,T))] \leq D_\mathrm{r}^\mathrm{f} \tag{7b}$$

in addition to the conditions in (6).

4. THE GAUSSIAN-QUADRATIC CASE

Consider a white Gaussian source with a white Gaussian reference channel. Specifically, we model the source as an i.i.d. Gaussian sequence where each S_i has mean zero and variance σ_S^2, and the independent reference channel noise as an i.i.d. sequence whose ith element N_i has mean zero and variance σ_N^2. Furthermore, we adopt the quadratic distortion measure $d(a,b) = (a-b)^2$.

A simple inner bound on the distortion region is obtained [1] by ignoring the authenticity requirement (1)

$$D_\mathrm{r} \geq \frac{\sigma_N^2 \sigma_S^2}{\sigma_N^2 + \left(\sqrt{D_\mathrm{e}} + \sigma_S\right)^2}. \tag{8}$$

To derive outer bounds we numerically optimize over all distributions where (S, U, X) are jointly Gaussian. For

[1] Although public key schemes are possible, to simplify and focus the exposition we consider only secret key schemes in this paper.

[2] When its argument is a set, $|\cdot|$ denotes its cardinality.

low D_{e}, we can closely approach the numerically obtained outer bound by using a distribution with structure similar to that used to achieve capacity in the related problem of information embedding [2]. For this encoding structure we let $U = S + T/\alpha$ and $X = U + (1-\alpha)(S-U) = S + T$ where T is a Gaussian random variable with mean zero and variance σ_T^2 independent of both the source S and the channel noise N.

The encoding distortion is simply $D_{\mathrm{e}} = \sigma_T^2$. The best reconstruction distortion is obtained by choosing $g(\cdot)$ to be the minimum mean-square estimate of S given U yielding

$$D_{\mathrm{r}} = E[S^2]\left(1 - \frac{E[SU]^2}{E[S^2]E[U^2]}\right) = \frac{\sigma_S^2 D_{\mathrm{e}}}{D_{\mathrm{e}} + \alpha^2\sigma_S^2}. \qquad (9)$$

The best choice of α is [1]

$$\alpha_{\text{auth}} = \alpha_{\text{ie}}\left(1 + \sqrt{1 + \frac{D_{\mathrm{e}} + \sigma_N^2}{\sigma_S^2}}\right)$$

where $\alpha_{\mathrm{ie}} = D_{\mathrm{e}}/(D_{\mathrm{e}} + \sigma_N^2)$ is the corresponding information embedding scaling parameter determined by Costa [2]. Evidently, the scaling parameter for the authentication problem is at least twice the scaling for information embedding and significantly larger when either the SNR σ_S^2/σ_N^2 or signal-to-distortion ratio (SDR) $\sigma_S^2/D_{\mathrm{e}}$ is small.

For high D_{e}, an encoder that essentially amplifies the source to overcome the reference channel noise closely approaches the numerically obtained inner bound. This structure corresponds to choosing the encoder random variables according to $U = S + T$ and $X = \beta U$. In turn, choosing as $g(\cdot)$ the minimum mean-square error estimator of S given U yields the distortions

$$D_{\mathrm{e}} = (1-\beta)^2\sigma_S^2 + \beta^2\sigma_T^2 \qquad \text{and} \qquad D_{\mathrm{r}} = \frac{\sigma_S^2\sigma_T^2}{\sigma_S^2 + \sigma_T^2},$$

where the choice

$$\beta = \sqrt{\frac{\sigma_S^2\sigma_N^2}{\sigma_T^2(\sigma_S^2 + \sigma_T^2)}} \qquad (10)$$

can be shown [1] to satisfy the mutual information constraint in (6a).

Figure 3 compares the distortion regions corresponding to the inner and outer bounds described above to a simple quantize-and-embed alternative strategy. In this strategy, which is at the heart of many previously proposed authentication schemes, the encoder compresses the source signal, protects it with cryptographic techniques and embeds the result in the original signal. The receiver decodes the embedded message, cryptographically verifies its authenticity, and reconstructs the compressed signal. The best possible performance of such quantize-and-embed systems is [1]

$$D_{\mathrm{r}} = \frac{\sigma_S^2\sigma_N^2}{\sigma_N^2 + D_{\mathrm{e}}},$$

which we can verify is a factor of up to $\mathrm{SNR}/2$ worse in reconstruction distortion than the optimum system [1], as the figure reflects.

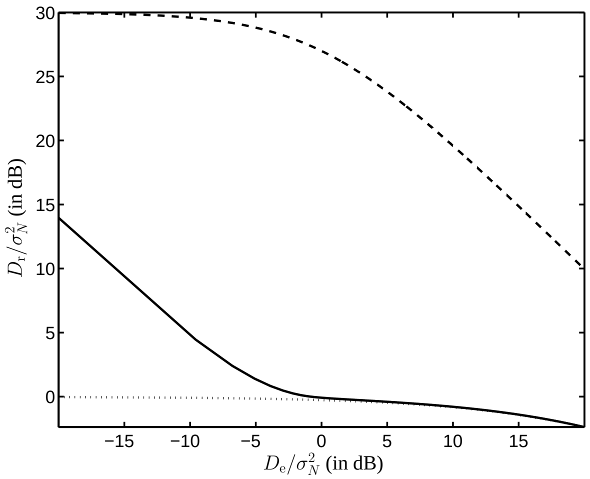

Fig. 3. Comparison of achievable distortion regions for the quantize-and-embed strategy (broken line), and inner (dotted line) and outer (solid line) bounds on the optimal distortion region obtained in [1] for the Gaussian-quadratic problem with SNR of 30 dB. When the reconstructions are not required to be authentic, all points above the dotted line are achievable: the security requirement strictly decreases the achievable distortion region.

A corresponding achievable region for layered systems in the Gaussian-quadratic scenario is developed in [1].

5. CONCLUDING REMARKS

While this paper describes an optimum architecture for multimedia authentication systems, many aspects of the detailed design and implementation of such systems remain to be addressed. Examples span a variety of fields and include information-theoretic issues such as error exponent behavior, communication-theoretic issues such as the design and analysis of structured authentication codes, signal processing considerations such as appropriate source and reference channel models and the choice of a good signal basis, and topics in computer science regarding cryptographic tools for public and private key versions of such systems.

6. REFERENCES

[1] E. Martinian, G. Wornell, and B. Chen, "Authentication with distortion criteria," submitted to *IEEE Trans. Inform. Theory*.

[2] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 3, pp. 439–441, May 1983.