AUDIO AUTHENTICATION BASED ON DISTRIBUTED SOURCE CODING

David Varodayan, Yao-Chung Lin and Bernd Girod

Information Systems Laboratory, Stanford University, Stanford, CA 94305
{varodayan, yao-chung.lin, bgirod}@stanford.edu

ABSTRACT

Audio authentication is important in content delivery via untrusted intermediaries, for example peer-to-peer (P2P) file sharing. Many differently encoded versions of the original audio might exist. Distinguishing the legitimate diversity of encodings from malicious tampering is the challenge addressed in this paper.

We develop an approach based on distributed source coding for the problem of backward-compatible audio authentication. The key idea is to provide a Slepian-Wolf encoded quantized perceptually significant audio projection as authentication data. This version can be correctly decoded only with the help of authentic audio as side information. Distributed source coding provides the desired robustness against legitimate encoding variations, while detecting illegitimate modification. We demonstrate reliable authentication at a Slepian-Wolf bitrate of less than 100 bit/s.

Index Terms— Audio authentication, distributed source coding, LDPC codes

1. INTRODUCTION

Media authentication is important in many applications of content delivery via untrusted intermediaries, such as peer-to-peer (P2P) file sharing. In this kind of application, many differently encoded versions of the original media file might exist. Moreover, transcoding or other acceptable modification at intermediate nodes might be required, giving rise to further diversity. On the other hand, intermediaries might tamper with the contents for a variety of reasons, such as interfering with the distribution of a particular file, piggybacking inauthentic content, or generally discrediting a particular distribution system. Distinguishing the legitimate diversity of encodings from malicious manipulation is the major technical challenge for media authentication systems. Past approaches fall into two groups: watermarks and media hashes.

A fragile watermark can be embedded into the host signal waveform without perceptual distortion [1]. Users can confirm the authenticity by extracting the watermark from the received content. The system design should ensure that the watermark survives lossy compression, but that it breaks as a result of a malicious manipulation. Unfortunately, watermarking authentication is not backward compatible with previously encoded contents; unmarked contents cannot be authenticated later. Embedded watermarks might also increase the bitrate required when compressing a media file.

Media hashing [2] achieves verification of previously encoded media by using an authentication server to supply authentication data to the user. Media hashes are inspired by cryptographic digital signatures [3], but unlike cryptographic hash functions, media hash functions are supposed to offer proof of perceptual integrity. Using a cryptographic hash, a single bit difference leads to an entirely different hash value. If two media signals are perceptually indistinguishable, they should have identical hash values. A common approach of media hashing is extracting features which have perceptual importance and should survive compression. The authentication data are generated by compressing these features or generating their hash values. The user checks the authenticity of the received content by comparing the features or their hash values to the authentication data.

In [4], we proposed an extension of hashing for image authentication based on distributed source coding. This paper adapts that architecture to audio authentication. Our approach has similarities with secure biometric authentication [5]. They are also related to the semi-fragile watermarking scheme for images in [6], which, however, is not applicable to authentication of previously encoded media.

In our audio authentication proposal, the authentication server provides a user with a Slepian-Wolf encoded audio projection, and the client attempts to decode this bitstream using the client's audio as side information. The Slepian-Wolf result [7] indicates that the lower the distortion between side information and the original, the fewer authentication bits are required for correct decoding. By correctly choosing the bitrate of the authentication data, this insight allows us to distinguish between legitimate encoding variations of the audio and illegitimate modifications. In Section 2, we describe the proposed audio authentication scheme and its rationale in detail. Simulation results are presented in Section 3.

2. AUDIO AUTHENTICATION SCHEME

Fig. 1 depicts our proposed audio authentication scheme. We denote the original source audio as A_o. The client receives the audio A_c as the output of a two-state lossy channel that models legitimate and illegitimate modifications. The left-hand side of Fig. 1 shows that the authentication data consist of a Slepian-Wolf encoded quantized audio projection of A_o and a digital signature of that version. The verification decoder, in the right-hand side of Fig. 1, knows the statistics of the worst permissible legitimate channel and can correctly decode the authentication data only with the help of authentic audio A_c as side information.

2.1. Two-State Channel

We model the client's audio A_c by way of a two-state lossy channel, shown in Fig. 2. In the legitimate state, the channel performs perceptual coding (such as mp3 or ogg) at a bitrate of 64 kbit/s or higher. In the illegitimate state, it additionally includes a malicious attack. We assume that A_o and A_c remain synchronized during these operations, and discuss this assumption in Section 3.

This work has been supported, in part, by the Max Planck Center for Visual Computing and Communication and, in part, by a gift from NXP Semiconductors to the Stanford Center for Integrated Systems.

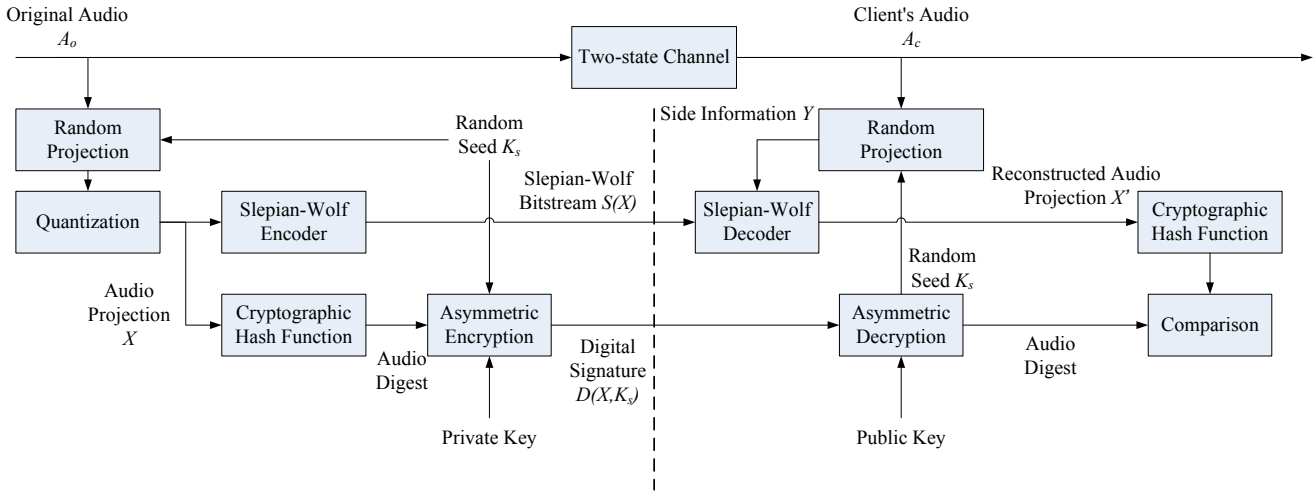

Fig. 1. Audio authentication system based on distributed source coding

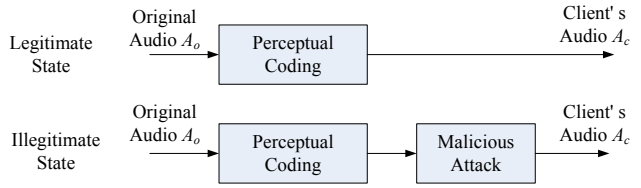

Fig. 2. Two-state lossy channel

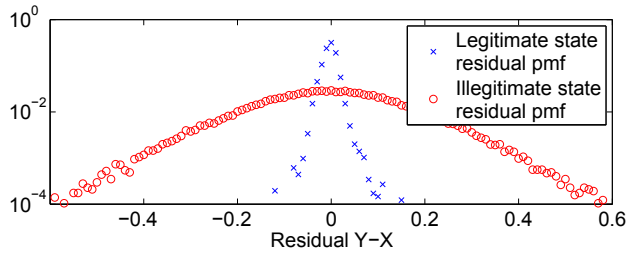

Fig. 3. The residual distributions between audio projections of channel input and output in legitimate and illegitimate states

To study the properties of this channel, we consider 8 audio tracks of duration about 3 s (65664 samples at 22.05 kHz) as instances of the original audio A_o. In the legitimate state, the client's audio A_c is generated by mp3-coding each track A_o at constant bitrate 64 kbit/s. In the illegitimate state, additional unauthorized content is inserted in three steps. First, the mp3-coded version of A_o is multiplied by an envelope $\frac{1}{2} + \frac{1}{2}\cos(\omega t)$, where ω is chosen so that the audio duration matches one oscillation period. Then another track is mp3-coded at constant bitrate 64 kbit/s and multiplied by an envelope $\frac{1}{2} - \frac{1}{2}\cos(\omega t)$. Finally, these two tracks are added together to produce the client's audio A_c.

The joint statistics of the input and output vary depending on the state of the channel. Fig. 3 compares the distribution of the residual $Y - X$, where X and Y are projections of the original audio A_o and the client's audio A_c, respectively. Each projection produces a sequence of pseudorandomly weighted coefficients of tonality [8], and will be described in Section 2.2. Fig. 3 shows that the legitimate state of the channel produces a much narrower residual distribution than the illegitimate state, and it is this difference in the joint statistics of X and Y that is exploited for authentication.

2.2. Authentication Data Generation

In our authentication system shown in Fig. 1, a pseudorandom projection (based on a randomly drawn seed K_s) is applied to the original audio A_o and the projection coefficients are quantized to yield X. The authentication data comprise two parts, both derived from X. The Slepian-Wolf bitstream $S(X)$ is the output of a Slepian-Wolf encoder based on low-density parity-check (LDPC) codes and the much smaller digital signature $D(X, K_s)$ consists of the seed K_s and a cryptographic hash value of X signed with a private key.

The authentication data are generated by a server upon request. Each response uses a different random seed K_s, which is provided to the decoder as part of the authentication data. This prevents an attack which simply confines the tampering to the nullspace of the projection.

The projection itself is inspired by the coefficient of tonality, one parameter in a perceptual model for audio [8]. First the audio is divided into overlapping frames of 256 samples with overlaps of 128 samples. Each frame is windowed sinusoidally and transformed by the MDCT to create 128 frequency coefficients. The magnitudes of the frequency coefficients are summed into bins $B[i]$ (for $i = 1, \ldots, 25$) that demarcate the 25 critical bands of human hearing. The perceptual spreading of energy is modeled by the function

$$F[i] = 10^{0.1\left(15.81+7.5(i+0.474)-17.5\sqrt{1+(i+0.474)^2}\right)},$$

so the apparent distribution of energy among the critical bands is

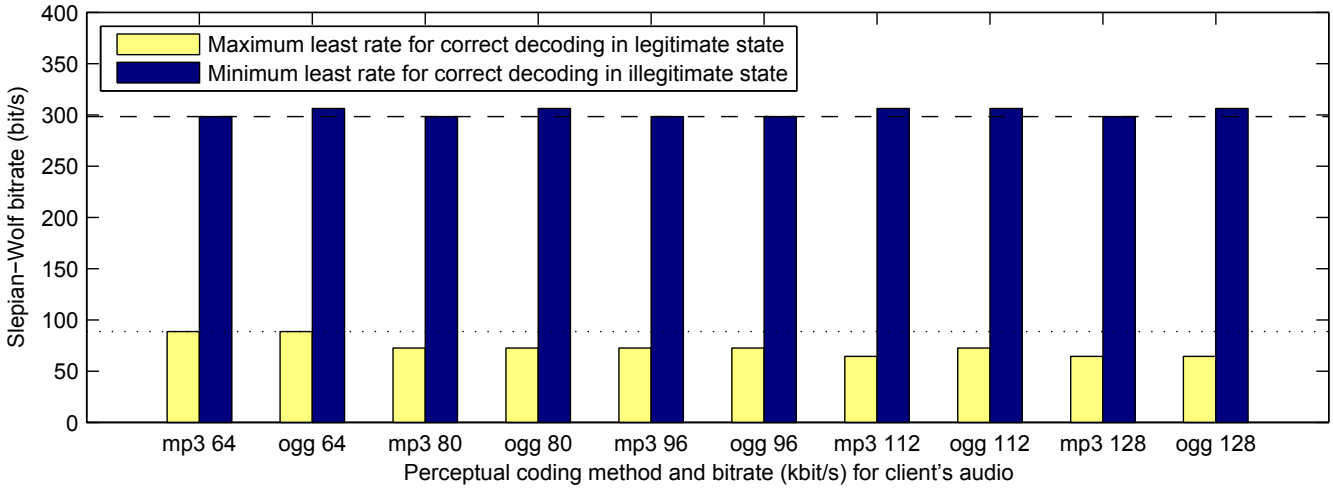

Fig. 4. Slepian-Wolf bitrates for 3 bits of quantization of X

given by convolution:

$$C[i] = \sum_{j=1}^{25} B[j]F[i-j], \text{for } i = 1, \ldots, 25.$$

Then a weighted arithmetic mean (AM) and geometric mean (GM) of C are computed:

$$\begin{aligned} \mu_a &= \sum_{i=1}^{25} C[i]w[i] \\ \mu_g &= \prod_{i=1}^{25} C[i]^{w[i]}, \end{aligned}$$

where $w[i]$ are pseudorandom weights (based on the random seed K_s), drawn independently from a Gaussian distribution $\mathcal{N}(1, \sigma^2)$ clipped at zero and normalized so that $\sum w[i] = 1$. We choose $\sigma = 0.2$ empirically. The ratio

$$\text{WSFM} = \frac{\mu_g}{\mu_a}$$

is the weighted spectral flatness measure, which is bounded between 0 and 1. The upper bound 1 is achieved if and only if $C[i]$ is constant for all i, according to the AM-GM inequality. Values near 1 correspond to a relatively flat spectrum, perceived as noisy, while values near 0 correspond to perception of the frame of audio as tone-like. Note that WSFM is also invariant to volume change. We then choose to define the coefficient of tonality α, for our own purposes, as

$$\alpha = \min\left(\frac{10\log_{10}\text{WSFM}}{-5}, 1\right),$$

which is also bounded between 0 and 1, but is inversely related to WSFM. Thus, the audio projection produces a sequence of coefficients of tonality, one per frame of audio and each built from different weights. The sequence is quantized to form X, a pseudorandom perceptually significant feature that is invariant to volume change.

The rate of the Slepian-Wolf bitstream $S(X)$ determines how statistically similar the client's audio must be to the original to be declared authentic. If the conditional entropy $H(X|Y)$ exceeds the bitrate R, then X can no longer be decoded correctly [7]. Therefore, the rate of $S(X)$ should be chosen to distinguish between the different joint statistics induced in the audio by the legitimate and illegitimate channel states. At the encoder, we select a Slepian-Wolf bitrate just sufficient to authenticate legitimate mp3 or ogg encodings of the original audio at constant bitrate 64 kbit/s.

2.3. Authentication Data Verification

At the receiver, the user seeks to authenticate the audio A_c with authentication data $S(X)$ and $D(X, K_s)$. It first projects A_c to Y in the same way as during authentication data generation. A Slepian-Wolf decoder reconstructs X' from the Slepian-Wolf bitstream $S(X)$ using Y as side information. Decoding is via LDPC belief propagation with a joint bitplane module [9] initialized according to the statistics of the legitimate channel state at the worst permissible quality for the given original audio. Finally, the audio digest of X' is computed and compared to the audio digest, decrypted from the digital signature $D(X, K_s)$ using a public key. If these two digests are identical, the receiver recognizes audio A_c as authentic.

3. AUTHENTICATION RESULTS

As described in Section 2.1, there are 8 original tracks A_o, each of duration about 3 s (65664 samples at 22.05 kHz). We operate the legitimate channel state as perceptual coding using mp3 or ogg at constant bitrates of 64, 80, 96, 112 and 128 kbit/s. The illegitimate channel multiplies the perceptually coded A_o by an envelope $\frac{1}{2} + \frac{1}{2}\cos(\omega t)$, and then adds another track, coded the same way as A_o and multiplied by an envelope $\frac{1}{2} + \frac{1}{2}\cos(\omega t)$. As mentioned in Section 2.1, ω is chosen so that the audio duration matches one oscillation period. In this way, the illegitimate channel models the insertion of unauthorized content.

The audio projection X consists of 512 coefficients of tonality, quantized to between 1 and 8 bits. The Slepian-Wolf codec is implemented using LDPC Accumulate (LDPCA) codes [10] with joint bitplane decoding [9]. For each combination of A_o and A_c at ev-

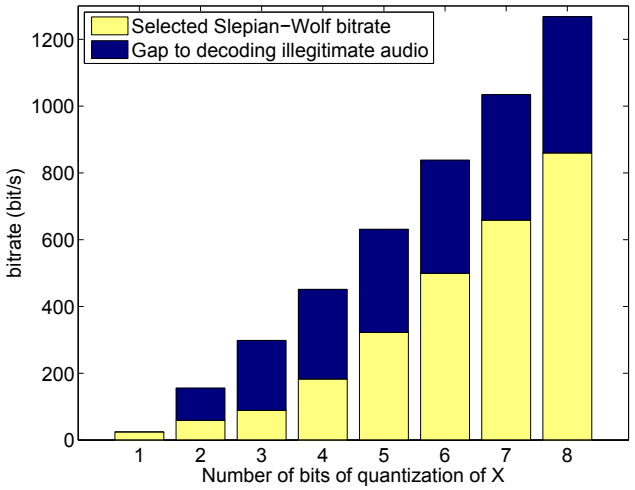

Fig. 5. Selected Slepian-Wolf bitrates (in bit/s) for different numbers of bits of quantization

ery encoding, we measure the least rate for correct decoding of the Slepian-Wolf coded audio projection $S(X)$.

Fig. 4 summarizes the data for 3 bits of quantization of X. For each perceptual coding method, a pale bar shows the least rate for decoding legitimate audio, maximized over all 8 tracks. The corresponding dark bar shows the least rate for decoding illegitimate audio, minimized over the 56 combinations of original and inserted tracks. The dotted line is the maximum level among pale bars, indicating the selected Slepian-Wolf bitrate for 3 bits of quantization. This rate of 89 bit/s is just sufficient to authenticate mp3 and ogg encodings at the lowest desired quality. The size of the gap to the dashed line (the minimum level among dark bars) justifies our claim that the Slepian-Wolf bitrate reliably distinguishes between the legitimate and illegitimate audio.

Fig. 5 plots the selected Slepian-Wolf bitrates and the gaps to the minimum rates for decoding illegitimate audio, as the number of bits of quantization varies. Observe that 1 bit of quantization requires a Slepian-Wolf rate of 24 bit/s, but this is insufficient to distinguish between legitimate and illegitimate channel states. In fact, the probability of false acceptance of inauthentic audio A_c is 40.5%. On the other hand, the selected Slepian-Wolf rates for all other levels of quantization yield zero probabilities of false acceptance. In particular, a Slepian-Wolf rate of 89 bit/s corresponds to 3 bits of quantization, which offers the largest rate gap relative to the Slepian-Wolf rate. Note that our choice for the Slepian-Wolf rate defines the probability of false rejection of authentic audio A_c to be zero.

We have so far assumed that the original audio A_o and the client's audio A_c remain synchronized. This assumption is unjustified; even in the legitimate channel state, the perceptual coding operation may delay the audio. In future work, we plan to incorporate unsupervised synchronization learning into the Slepian-Wolf decoder using an Expectation Maximization algorithm [11]. We have already demonstrated this method for joint distributed source decoding and parameter learning in the context of stereo image compression [9, 12, 13].

4. CONCLUSION

In this work, we developed a backward-compatible audio authentication scheme, based on distributed source coding, that distinguishes between legitimate encoding variations of audio and illegitimately modified versions. Our system uses a pseudorandom audio projection that is perceptually significant. A Slepian-Wolf bitrate of less than 100 bit/s is demonstrated to be sufficient for reliable authentication. At this rate, a single IP packet sent by the server would suffice to verify the integrity of 2 minutes of audio.

5. REFERENCES

[1] M. Steinebach and J. Dittmann, "Watermarking-based digital audio data authentication," *EURASIP J. Applied Signal Proc.*, vol. 2003, no. 10, pp. 1001–1015, 2003.

[2] J. Haitsma, T. Kalker, and J. Oostveen, "Robust audio hashing for content identification," in *Proc. Internat. Workshop Content-Based Multimedia Indexing*, Brescia, Italy, 2001.

[3] W. Diffie and M. E. Hellman, "New directions in cryptography," *IEEE Trans. Inform. Theory*, vol. 22, no. 6, pp. 644–654, Jan. 1976.

[4] Y.-C. Lin, D. Varodayan, and B. Girod, "Image authentication based on distributed source coding," in *Proc. IEEE Internat. Conf. Image Processing*, San Antonio, TX, 2007.

[5] E. Martinian, S. Yekhanin, and J. Yedidia, "Secure biometrics via syndromes," in *Proc. Allerton Conf. Commun., Contr. and Comput.*, Allerton, IL, 2005.

[6] Q. Sun, S.-F. Chang, M. Kurato, and M. Suto, "A new semi-fragile image authentication framework combining ECC and PKI infrastructure," in *Proc. IEEE Internat. Symp. Circuits and Syst.*, Phoenix, AZ, 2002.

[7] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. 19, no. 4, pp. 471–480, July 1973.

[8] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–513, Apr. 2000.

[9] D. Varodayan, A. Mavlankar, M. Flierl, and B. Girod, "Distributed grayscale stereo image coding with unsupervised learning of disparity," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, 2007.

[10] D. Varodayan, A. Aaron, and B. Girod, "Rate-adaptive distributed source coding using low-density parity-check codes," in *Proc. Asilomar Conf. on Signals, Syst., Comput.*, Pacific Grove, CA, 2005.

[11] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc., Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[12] D. Varodayan, A. Mavlankar, M. Flierl, and B. Girod, "Distributed coding of random dot stereograms with unsupervised learning of disparity," in *Proc. IEEE Internat. Workshop Multimedia Signal Processing*, Victoria, BC, Canada, 2006.

[13] D. Varodayan, Y.-C. Lin, A. Mavlankar, M. Flierl, and B. Girod, "Wyner-Ziv coding of stereo images with unsupervised learning of disparity," in *Proc. Picture Coding Symp.*, Lisbon, Portugal, 2007.