Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx-08), Espoo, Finland, September 1-4, 2008

DETECTION AND IDENTIFICATION OF SPARSE AUDIO TAMPERING USING DISTRIBUTED SOURCE CODING AND COMPRESSIVE SENSING TECHNIQUES

Giorgio Prandi
Dipartimento di Elettronica e Informazione,
Politecnico di Milano, Italy
prandi@elet.polimi.it

Giuseppe Valenzise
Dipartimento di Elettronica e Informazione,
Politecnico di Milano, Italy
valenzise@elet.polimi.it

Marco Tagliasacchi
Dipartimento di Elettronica e Informazione
Politecnico di Milano, Italy
tagliasa@elet.polimi.it

Augusto Sarti
Dipartimento di Elettronica e Informazione
Politecnico di Milano, Italy
sarti@elet.polimi.it

ABSTRACT

In most practical applications, for the sake of information integrity not only it is useful to detect whether a multimedia content has been modified or not, but also to identify which kind of attack has been carried out. In the case of audio streams, for example, it may be useful to localize the tamper in the time and/or frequency domain. In this paper we devise a hash-based tampering detection and localization system exploiting compressive sensing principles. The multimedia content provider produces a small hash signature using a limited number of random projections of a time-frequency representation of the original audio stream. At the content user side, the hash signature is used to estimate the distortion between the original and the received stream and, provided that the tamper is sufficiently sparse or sparsifiable in some orthonormal basis expansion or redundant dictionary (e.g. DCT or wavelet), to identify the time-frequency portion of the stream that has been manipulated. In order to keep the hash length small, the algorithm exploits distributed source coding techniques.

1. INTRODUCTION

The delivery of multimedia contents in peer-to-peer networks might give rise to different versions of the same multimedia object at different nodes. In the case of audio files, some versions might differ from the original because of processing due, for instance, to transcoding or bitstream truncation. In other cases, malicious attacks might occur by tampering part of the audio stream and possibly affecting its semantic content. In this paper we propose to add a small hash to the audio stream to detect and identify audio tampers. At the content user side, the information contained in the hash enables to estimate the distortion of the received audio stream with respect to the original version, and to localize the tampering, if the attack is sparse in one of the analyzed basis expansion.

In the past few years, hashes have been used for the protection of multimedia contents, especially in the field of image authentication. In [1], the authors propose a system that performs image authentication using distributed source codes. The hash consists of syndrome bits applied to quantized random projections of the original image. To perform authentication, a Slepian-Wolf decoder receives an input the hash and the (possibly tampered) image, which serves as side information. If decoding succeeds, the image is declared authentic. The work has been extended in [2] to perform tampering localization, at the cost of extra syndrome bits. The authors of [3] present an interesting algorithm that produces hashes robust to some legitimate image manipulations (like cropping, scaling, rotation) and sensitive to illegal manipulations (like image tampering).

Watermarking has been used to solve the problem of tampering localization [4][5]. A fragile watermark is inserted into the image when it is created, and extracted during the authentication phase. Tampering can be localized by identifying the damage to the watermark. Watermarking techniques have also been used to solve the problem of content authentication and tampering localization in audio data. In [6] two complementary watermarks are embedded in audio signals to achieve audio protection and tampering localization. However, to the authors' knowledge, in the literature there are no previous works addressing the problem of identifying tampering in audio streams. In general, watermarking based schemes suffer from the following disadvantages: 1) watermarking authentication is not backward compatible with previously encoded contents (unmarked contents cannot be authenticated later by just retrieving the corresponding hash); 2) the original content is distorted by the watermark; 3) the bit-rate required to compress a multimedia content might increase due to the embedded watermark. Conversely, content hashing embeds a signature of the original content as part of the header information, or can provide a hash separately from the content upon a user's request. In order to limit the rate overhead, the size of the hash needs to be as small as possible. At the same time, the goal of tampering localization calls for increasing the hash size, in order to capture as much as possible about the original multimedia object. In this paper we explicitly target these conflicting requirements by proposing a hashing technique based on compressive sensing principles. The key tenet is that, if the tampering is sparse enough (or it can be sparsified in some orthonormal basis or redundant dictionary), it can be identified by means of a limited number of random projections of the original signal. In addition, in order to keep the size of the hash as small as possible, the hash information is encoded by exploiting distributed source coding tools.

The rest of the paper is organized as follows. Section 2 provides background information about compressive sensing and distributed source coding; Section 3 gives a detailed description of

the system. The results of tampering estimation are depicted in Section 4. Finally, Section 5 gives some concluding remarks.

2. BACKGROUND

2.1. Compressive sensing

In our system, compressive sensing principles are used to build the hash signature of the audio stream. Compressive sensing allows to capture and represent signals at rates below the Nyquist frequency [7]. In fact, it has been proved that it is possible to reconstruct a signal using a limited number of non-adaptive linear random projections that preserve the original structure of the signal, by solving a specific optimization problem. The main constraint which must be satisfied is that the signal has to be *sparse* or, at least, *compressible*, i.e. the signal can be represented in some basis expansion using only a few large magnitude coefficients. A more detailed discussion about compressive sensing can be found in [7].

2.2. Distributed source coding

As mentioned in previous sections, in our system we use a distributed source coding technique to reconstruct the hash signature of the audio stream at the content user side. Distributed source coding has been widely applied to video coding [8] in order to move the computational complexity from the encoder to the decoder side. According to distributed source coding principles stated by the Wyner-Ziv theorem, it is possible to perform lossy encoding with side information at the decoder. The side information represents a distorted version of the source, which is made available at the decoder side only. In our approach, the original information is the hash computed from the content provider, and the side information consists of the hash signature computed from the audio stream received at the user side (which may be modified with respect to the original). By requesting syndrome bits from the encoder, the decoder is able to correct the possibly distorted side information and to finally reconstruct the original hash signature. The more the side information is distorted, the more syndrome bits are needed to reconstruct the original hash; if the number of requested bits exceeds some pre-specified threshold, we may consider the received stream too distorted and completely unauthentic. Under normal conditions, the number of requested syndrome bits is significantly less than the number of bits of the original information, so, the hash reconstruction approach based on distributed source coding technique allows to save bits with respect to the direct transmission of the original hash from the content provider to the user.

3. DESCRIPTION OF THE SYSTEM

The proposed tampering detection and localization scheme is depicted in Figure 1. The producer of the original audio stream builds a small hash signature of the audio signal $\mathbf{X} \in \mathbb{R}^N$, where N is the total number of audio samples of the signal. The audio content can be distributed over a network through untrusted nodes where a modification of the signal may occur. The user receives the audio stream $\widetilde{\mathbf{X}}$. In order to perform content authentication, the user sends a request for the hash signature to the content provider. By exploiting the hash, the user can estimate the distortion of the received content $\widetilde{\mathbf{X}}$ with respect to the original $\mathbf{X}$. Furthermore, if the tampering is sparse in some basis expansion, the system produces a tamper estimation $\hat{\mathbf{e}}$ which identifies the attack in the time-frequency domain. At the content producer side, the encoder generates the hash signature $\mathcal{H}(\mathbf{X}, S)$ as follows:

1. *Frame based subband log-energy extraction*: the original single-channel audio stream $\mathbf{X}$ is partitioned into non-overlapping frames of size F. The power spectrum of each frame is subdivided into U Mel frequency subbands, and for each subband the related spectral log-energy is extracted. Let denote $h_{f,u}$ the energy value for the u-th band at frame f. The corresponding log-energy value is computed as follows:
$$x_{f,u} = \log\left(1 + h_{f,u}\right) \quad (1)$$
The log-energy values are stored in a vector $\mathbf{x} \in \mathbb{R}^n$, where $n = UN/F$ is the number of log-energy values extracted from the audio stream. Note that using Mel frequency subbands and log-energy values gives an immediate perceptual semantics to tamper detection and identification.
2. *Random projections*: A number of linear random projections $\mathbf{y} \in \mathbb{R}^m$, $m < n$, is produced as $\mathbf{y} = \mathbf{A}\mathbf{x}$. The entries of the matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ are sampled from a Gaussian distribution $\mathcal{N}(0, 1/n)$, using some random seed S, which will be sent as part of the hash to the user.
3. *Wyner-Ziv encoding*: The random projections $\mathbf{y}$ are quantized with a uniform scalar quantizer with step size Δ. Bitplane extraction is performed on the quantization bin indexes. Each bitplane is encoded by sending syndrome bits generated by means of a Low-Density Parity-Check Code (LDPC). The rate allocated to the hash depends on the expected distortion between the original and the tampered audio stream.

The content user receives the (possibly tampered) audio stream $\widetilde{\mathbf{X}}$ and requests the syndrome bits and the random seed of the hash $\mathcal{H}(\mathbf{X}, S)$ to the authentication server at content producer side. On each user's request, a different seed S is used in order to avoid that a malicious attack could exploit the knowledge of the nullspace of $\mathbf{A}$.

1. *Frame-based subband log-energy extraction*: computed on signal $\widetilde{\mathbf{X}}$ using the same algorithm described above for the content producer side. At this step, the vector $\widetilde{\mathbf{x}}$ is produced.
2. *Random projections*: $\widetilde{\mathbf{y}} = \mathbf{A}\widetilde{\mathbf{x}}$.
3. *Wyner-Ziv decoding*: A quantized version $\hat{\mathbf{y}}$ is obtained using the hash syndrome bits and $\widetilde{\mathbf{y}}$ as side information. LDPC decoding is performed starting from the most significant bitplane. If the actual distortion between the original and the tampered audio stream is higher than the maximum distortion expected by the original content producer (determined by the rate allocated to the hash signature) decoding might fail. In this case, the audio stream is declared to be completely unauthentic and no tampering localization can be provided.
4. *Distortion estimation*: If Wyner-Ziv decoding succeeds, an estimate of the distortion in terms of a perceptual SNR measure is computed using the projections of the subsampled energy spectrum of the tamper. Let $\hat{\mathbf{b}} = \hat{\mathbf{y}} - \widetilde{\mathbf{y}}$ the projections of the subsampled energy spectrum of the tamper; we compute an estimate of the perceptual SNR of the received audio stream as
$$SNR_P = 10 \log_{10} \frac{\|\hat{\mathbf{y}}\|_2^2}{\|\hat{\mathbf{b}}\|_2^2} \quad \text{[dB]} \quad (2)$$

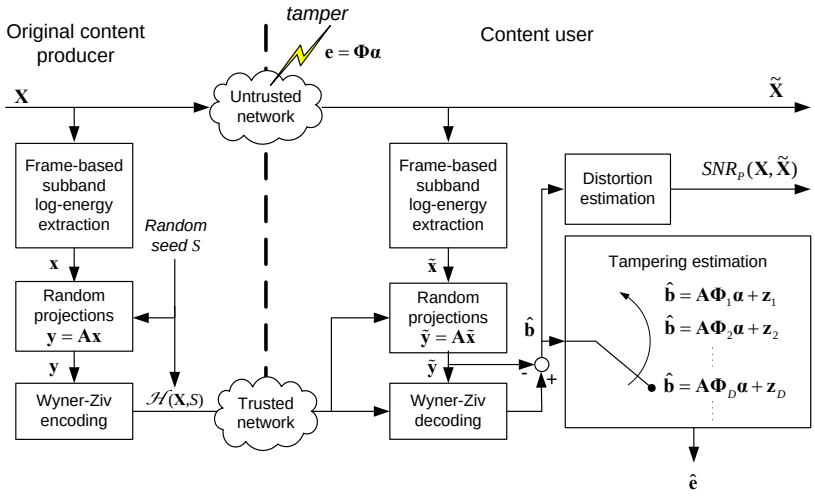


Figure 1: Block diagram of the proposed tampering localization scheme

5. *Tampering estimation*: An estimate of the tampering $\mathbf{e} = \tilde{\mathbf{x}} - \mathbf{x}$ can be obtained by solving the following undetermined system of linear equations:

$$\begin{aligned} \hat{\mathbf{b}} &= \hat{\mathbf{y}} - \tilde{\mathbf{y}} = \\ &= \mathbf{A}(\mathbf{x} - \tilde{\mathbf{x}}) + \mathbf{z} = \\ &= \mathbf{A}\mathbf{e} + \mathbf{z} \end{aligned} \tag{3}$$

where $\mathbf{z}$ is the hash quantization noise.

There exists an infinite number of solutions to (3); however, in the hypothesis that $\mathbf{e}$ is sparse, the optimal way for recovering e is to seek the sparsest solution of (3), i.e. the one that minimizes $\|\mathbf{e}\|_0$, where the ℓ_0 norm $\|\cdot\|_0$ simply counts the number of nonzeros entries of $\mathbf{e}$ [9]. Unfortunately, such a problem is NP hard and it is difficult to solve in practice. Nonetheless, recent literature about compressive sensing [9] has shown that, if $\mathbf{e}$ is sufficiently sparse, an approximation of it can be recovered by solving the following ℓ_1 minimization problem:

$$\hat{\mathbf{e}} = \min \|\mathbf{e}\|_1 \qquad \text{s.t.} \qquad \|\hat{\mathbf{b}} - \mathbf{A}\mathbf{e}\|_2 \leq \epsilon \tag{4}$$

where ϵ is chosen such that $\|\mathbf{z}\|_2 \leq \epsilon$. Problem (4) is a special instance of a second order cone program (SOCP) [9] and can be solved in $O(n^3)$ time. Nevertheless, several fast algorithms have been proposed in the literature that attempt to find the sparsest $\mathbf{e}$ satisfying the constraint $\|\mathbf{b} - \mathbf{A}\mathbf{e}\|_2 \leq \epsilon$. In our experiments, we adopt the SPGL1 algorithm [10], which is specifically designed for large scale sparse reconstruction problems. If $\mathbf{e}$ is not sparse enough with respect to the number of projections m, the solution found does not fulfil the constraint. In such cases, it is not possible to perform tampering localization in the original log-energy domain. However, it is possible to perform the analysis in other domains in which the tamper may be sparse. In our scheme, we assume that the tamper is sparse in some orthonormal basis $\boldsymbol{\Phi}$, so that:

$$\mathbf{e} = \boldsymbol{\Phi}\boldsymbol{\alpha} \tag{5}$$

where $\boldsymbol{\alpha}$ are the coefficients of the expansion of $\mathbf{e}$ in the basis $\boldsymbol{\Phi}$. In this case, instead of equation (3) we use the following one:

$$\hat{\mathbf{b}} = \mathbf{A}\boldsymbol{\Phi}_D\boldsymbol{\alpha} + \mathbf{z}_D \tag{6}$$

Due to the missing knowledge of basis $\boldsymbol{\Phi}$, one can try to sparsify the tamper in different bases $\boldsymbol{\Phi}_D$. In our system we have implemented the expansion of the tamper in the DCT, DCT 2D and Haar wavelet bases. If the tampering in the basis $\boldsymbol{\Phi}_D$ is sufficiently sparse, finding the minimum ℓ_1-norm solution of (6) allows to obtain a tampering estimation $\hat{\boldsymbol{\alpha}}$. Then, we can transform back the result to the original log-energy domain:

$$\hat{\mathbf{e}} = \Phi_D\hat{\boldsymbol{\alpha}} \tag{7}$$

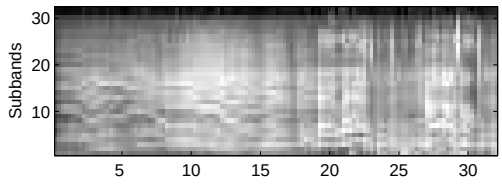
4. EXPERIMENTAL RESULTS

We have carried out some experiments on the first 32 seconds of Etta James' song "At last", sampled at 44100 Hz, at 16-bit per sample. The size of the audio frame has been set to $F = 11025$ samples (0.25 seconds), and the number of Mel frequency bands has been set to $U = 32$, obtaining a total of 128 audio frames corresponding to $n = 4096$ log-energy coefficients. The testbed has been built considering 3 kinds of tampering:

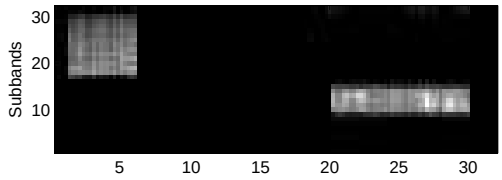
- *Time localized tampering (T)*: a time-limited audio fragment, taken from another portion of the whole song, is mixed with the original audio stream;
- *Frequency localized tampering (F)*: a low-pass phone-band filter (cut frequency at 3400 Hz and stop frequency at 4000 Hz) is applied to the entire original audio stream;
- *Time-frequency localized tampering (TF)*: a low-pass and a band-stop filters are applied to two different portions of the original audio stream (see Figure 2(b)).

We evaluate the goodness of the tampering estimation by calculating the normalized MSE between the log-energy spectrum of the original tamper and the log-energy spectrum of the estimated one:

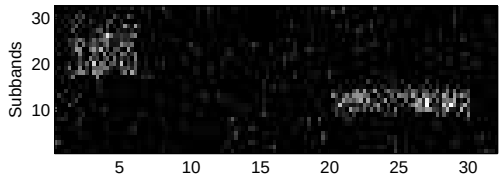
$$MSE_N = \frac{\|\hat{\mathbf{e}} - \mathbf{e}\|_2^2}{\|\mathbf{e}\|_2^2} \tag{8}$$



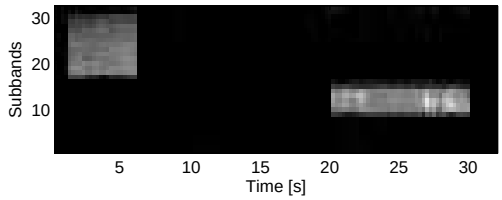
(a) Log-energy spectrum of the original audio signal



(b) Log-energy spectrum of the tamper



(c) Reconstructed tamper in log-energy domain. In this case the estimation reaches a Normalized MSE of $6.52 \cdot 10^{-2}$



(d) Reconstructed tamper in Haar wavelet domain. The Normalized MSE value is $3.01 \cdot 10^{-3}$

Figure 2: Example of time-frequency tampering consisting of a low-pass filter and a stop-band filter. The reconstruction has been performed using a 200 bps hash signature.

Results related to Normalized MSE obtained with a fixed bit rate for the hash are shown in Tables 1 (for 200 bps) and 2 (for 400 bps). From the tables it is clear that, by looking for a sparse tamper in other bases besides the canonical one (log-energy), better results can be achieved using the same hash length, as highlighted by the bold numbers in the tables. More precisely, for the time localized tampering the basis that gives the lower reconstruction distortion is the DCT one; for the frequency localized tampering, instead, the best basis w.r.t. reconstruction distortion is the DCT 2D. The time-frequency localized tampering is better reconstructed using the Haar Wavelet basis.

5. CONCLUSIONS

In this paper, a novel algorithm to detect and identify audio tampering by means of the recent compressive sensing framework has been described. Using the distributed source coding paradigm, we have shown how to produce very small yet effective hash signatures which allow to reliably identify tampering with a small information overhead; in addition, looking for a sparse representation of the tampering in some transformed domain enables to perform the identification with a further gain in hash payload overhead.

	Log-energy	DCT	DCT 2D	Haar Wavelet
T	$1.78 \cdot 10^{-2}$	$\mathbf{5.03 \cdot 10^{-4}}$	$2.88 \cdot 10^{-3}$	$8.22 \cdot 10^{-4}$
F	$7.80 \cdot 10^{-2}$	$5.57 \cdot 10^{-2}$	$\mathbf{2.88 \cdot 10^{-3}}$	$4.95 \cdot 10^{-3}$
TF	$6.52 \cdot 10^{-2}$	$4.78 \cdot 10^{-2}$	$1.09 \cdot 10^{-2}$	$\mathbf{3.01 \cdot 10^{-3}}$

Table 1: Distortion of tamper estimation MSE_N using a fixed bit rate for the hash signature of 200 bps.

	Log-energy	DCT	DCT 2D	Haar Wavelet
T	$1.27 \cdot 10^{-3}$	$\mathbf{4.27 \cdot 10^{-5}}$	$1.42 \cdot 10^{-3}$	$5.95 \cdot 10^{-5}$
F	$6.71 \cdot 10^{-2}$	$3.23 \cdot 10^{-2}$	$\mathbf{1.22 \cdot 10^{-3}}$	$1.95 \cdot 10^{-3}$
TF	$6.47 \cdot 10^{-3}$	$1.20 \cdot 10^{-2}$	$4.84 \cdot 10^{-3}$	$\mathbf{2.19 \cdot 10^{-4}}$

Table 2: Distortion of tamper estimation MSE_N using a fixed bit rate for the hash signature of 400 bps.

6. REFERENCES

[1] Y.C. Lin, D. Varodayan, and B. Girod, "Image authentication based on distributed source coding," in *IEEE International Conference on Image Processing*, S.Antonio, TX, Sept. 2007, vol. 3.

[2] Y.C. Lin, D. Varodayan, and B. Girod, "Spatial Models for Localization of Image Tampering Using Distributed Source Codes," in *Picture Coding Symposium (PCS)*, Lisbon, Portugal, Nov. 2007.

[3] S. Roy and Q. Sun, "Robust Hash for Detecting and Localizing Image Tampering," in *IEEE International Conference on Image Processing*, S.Antonio, TX, 2007, vol. 6.

[4] J. Fridrich, "Image watermarking for tamper detection," in *IEEE International Conference on Image Processing*, Chicago, Oct. 1998, vol. 2.

[5] J.J. Eggers and B. Girod, "Blind watermarking applied to image authentication," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, 2001, vol. 3.

[6] C.S. Lu, H.Y.M. Liao, and L.H. Chen, "Multipurpose audio watermarking," in *Proc. 15th Int. Conf. on Pattern Recognition*, 2000.

[7] R.G. Baraniuk, "Compressive Sensing," *Signal Processing Magazine, IEEE*, vol. 24, no. 4, pp. 118–121, 2007.

[8] B. Girod, AM Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 71–83, 2005.

[9] E. Candes, "Compressive sampling," in *International Congress of Mathematicians*, Madrid, Spain, 2006.

[10] E. van den Berg and M. P. Friedlander, "In pursuit of a root," Tech. Rep. TR-2007-19, Department of Computer Science, University of British Columbia, June 2007, Preprint available at `http://www.optimization-online.org/DB_HTML/2007/06/1708.html`.