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

G. Prandi, G. Valenzise, M. Tagliasacchi, A. Sarti

Dipartimento di Elettronica e Informazione - Politecnico di Milano

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. Background

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].
- 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.
- The signal has to be sparse or compressible (it can be represented in some basis expansion using only a few large magnitude coefficients).

Distributed source coding technique, widely applied to video coding [8], is used to reconstruct the hash signature of the audio stream at the content user side:

- 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. 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.
- In normal conditions, 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. Experimental Results

We carried out some experiments on the first 32 seconds of Etta James' song "At last" (sampled at 44100 Hz, 16-bit per sample).
We set the lenght of each frame F = 11025 samples (0.25 seconds), and the number of Mel frequency bands U = 32;
We have a total of 128 audio frames and n = 4096 log-energy coefficients.

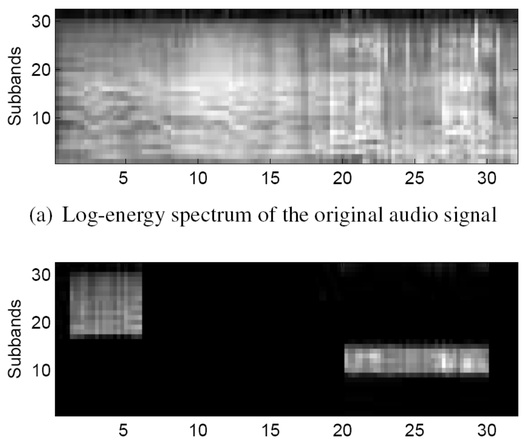
3 kinds of tampering applied to the original audio stream:

- *Time localized tampering (T)*: a time-limited audio fragment is mixed with the original audio stream;
- *Frequency localized tampering (F)*: a low-pass phone-band filter is applied to the entire 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 *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:

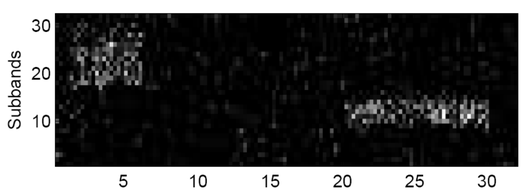
$$MSE_N = \frac{\|\hat{\mathbf{e}} - \mathbf{e}\|_2^2}{\|\mathbf{e}\|_2^2}$$

Results obtained using fixed bit rates for the hash (200 and 400 bps) are shown in Tables 1 and 2:

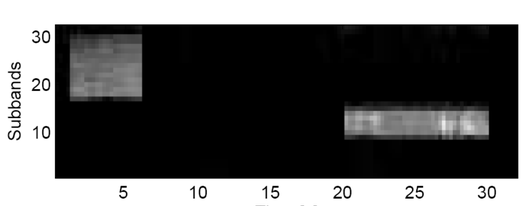


(a) Log-energy spectrum of the original audio signal

(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 $3.01 \cdot 10^{-3}$

	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	$6.52 \cdot 10^{-2}$	$4.78 \cdot 10^{-2}$	$1.09 \cdot 10^{-2}$	$\mathbf{3.01 \cdot 10^{-3}}$

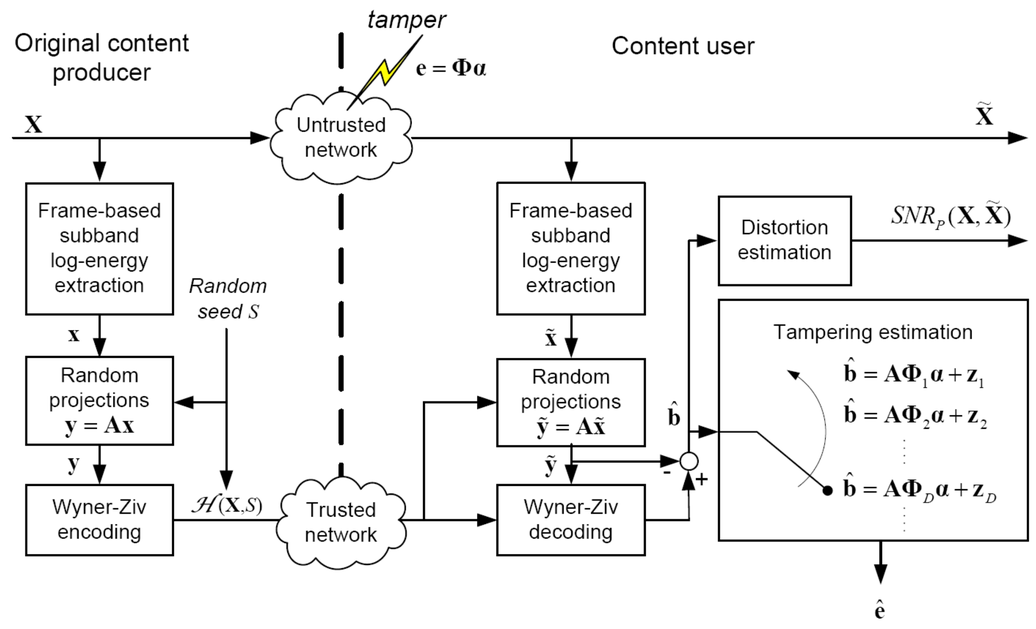
Table 1: Distortion of tamper estimation MSE_N using a fixed bit rate for the hash signature of 200 bps.

	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.

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.

2. Description of the system



Original Content Producer side

Frame based subband log-energy extraction

- The power spectrum of each non-overlapping audio frame of length *F* is subdivided into *U* Mel frequency subbands;
- For each subband the related spectral log-energy is extracted, producing a global vector **x** of *n* log-energy values.

Random projections

- A number of linear random projections from the vector of log-energy values is produced as $\mathbf{y} = \mathbf{A}\mathbf{x}$. ($\mathbf{A} \in \Re^{m}$, $\mathbf{y} \in \Re^{m}, m < n$);
- The entries of the matrix **A** are sampled from a Gaussian distribution $N\left(0, \frac{1}{n}\right)$, using some random seed S, which will be sent as part of the hash to the user.

Wyner-Ziv encoding

- The random projections $\mathbf{y}$ are quantized with a uniform scalar quantizer;
- Bitplane extraction is performed on the quantization bin indexes. Syndrome bits are 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.

Content User side

Frame-based subband log-energy extraction
Computed on signal $\tilde{\mathbf{X}}$ using the same algorithm described above for the content producer side. At this step, the vector $\tilde{\mathbf{x}}$ is produced.

Random projections:
$\tilde{\mathbf{y}} = \mathbf{A}\tilde{\mathbf{x}}$

Wyner-Ziv decoding

- A quantized version $\hat{\mathbf{y}}$ of $\mathbf{y}$ is obtained using the hash syndrome bits and $\tilde{\mathbf{y}}$ as side information;
- If the actual distortion between the original and the tampered audio stream is higher than the maximum distortion expected by the original content producer, the audio stream is declared to be completely unauthentic and no tampering localization can be provided.

Distortion estimation (perceptual SNR of the received audio stream): $SNR_P = 10\log_{10}\frac{\|\hat{\mathbf{y}}\|_2^2}{\|\hat{\mathbf{b}}\|_2^2}$ [dB]

Tampering estimation

- An estimate of the tampering $\mathbf{e} = \mathbf{x} - \tilde{\mathbf{x}}$ can be obtained by solving the following undetermined system of linear equations (**z** is the quantization noise): $\hat{\mathbf{b}} = \hat{\mathbf{y}} - \tilde{\mathbf{y}} = \mathbf{A}(\mathbf{x} - \tilde{\mathbf{x}}) + \mathbf{z} = \mathbf{A}\mathbf{e} + \mathbf{z}$
- The optimal way for recovering **e** is to seek the sparsest solution of the system (ℓ_0 norm). Unfortunately, the problem is NP hard. However, if **e** is sufficiently sparse, an approximation of **e** can be recovered by solving the following problem [9]: $\hat{\mathbf{e}} = \min\|\mathbf{e}\|_1 \quad \text{s.t.} \quad \left\|\hat{\mathbf{b}} - \mathbf{A}\mathbf{e}\right\|_2 \le \varepsilon$
- If the error **e** is not sufficiently sparse, we can try to find the solution in other domains (DCT, DCT 2D and Haar wavelet) by defining the following modified linear system: $\hat{\mathbf{b}} = \mathbf{A}\mathbf{\Phi}_D\boldsymbol{\alpha} + \mathbf{z}_D$
- In our scheme, we assume that the tamper is sparse in some orthonormal basis $\mathbf{\Phi}$, which is unknown. When an estimate of $\boldsymbol{\alpha}$ is computed, we can transform back the result to the original log-energy domain.

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.