Proceedings of the International Symposium on Musical Acoustics, March 31st to April 3rd 2004 (ISMA2004), Nara, Japan

Audio Watermarking Based on Sinusoidal Analysis and Synthesis

Yi-Wen Liu, Julius O. Smith

Center for Computer Research in Music and Acoustics, Stanford University, USA
{jacobliu,jos}@ccrma.stanford.edu

Abstract

We present a non-canonical approach to audio watermarking based on the introduction of small frequency changes to sinusoids. This idea is enabled by the sine + transient + residual signal decomposition. In this paper, we describe the signal decomposition in detail, and provide a procedure to embed and decode watermarks accordingly. To hide 43 bits/sec of information in an instrumental recording, experiments are conducted, and implications of the results are discussed.

1. Introduction

Audio watermarking refers to hiding binary information in sounds. Among existing applications such as copyright protection and broadcast monitoring, audio watermarks are required to be *transparent* and *robust*. Here, "transparency" means that the watermarks introduce no audible distortion, and "robustness" means that the watermarks can not be erased without degrading the perceptual quality of sounds. Satisfying these requirements, a watermark embedder is designed to hide as much information as possible.

Currently, most audio watermarking methods utilize the masking phenomena of human hearing. Transform coefficients of a signal are amplitude-modulated to carry information (e.g. [1][2]). The modulation is small enough to introduce no distortion above the masking threshold, but large enough so that any further distortion is audible. Instead, Our method utilizes the human just noticeable difference (JND) in frequency. The magnitude of frequency modulation is within JND, and watermark enbedding and decoding procedures are designed to be robust to reasonable signal manipulations, such as the MP3 compression.

The watermark embedding and decoding procedures are presented in Sec. 2 and 3, respectively. Selected experiments are described in Sec. 4, with discussions of the results in Sec. 5. Conclusions are given in Sec. 6.

2. System overview: watermark embedding

As mentioned in Sec. 1, our system embeds information by small frequency modulations. For this to happen, we have to first extract a sinusoidal portion of a signal. This process is called *sinusoidal analysis* here. In the sinusoidal analysis, rapid transient regions are labeled in time, and the remaining of a signal is decomposed into the sinusoidal and the residual parts in frequency [3]. In the current implementation, watermarks are only embedded to the sinusoidal part. The transient and the residual parts are kept unchanged. Afterwards, all three parts are superposed to form the watermarked signal. A block diagram of the analysis is shown in Fig. 1, and details are described in the following subsections.

2.1. Windowing and Fourier transform

In the analysis, we choose to use the Blackman-Harris window [4] of length 2048 for short time Fourier transform (STFT)[5], assuming a sampling rate of 44.1kHz. The windowed STFT is taken every 512 samples, which is the maximum hop-size for perfect reconstruction.

2.2. Transient detection

The reason for detecting transients is to avoid *pre-echos*. A pre-echo often accompanies a rapid attack when the short-time phase spectrum is modified. Since a rapid attack is often preceded by silence originally, a pre-echo is heard as an artifact. To prevent pre-echos, we detect rapid transient regions and keep them unchanged.

The transient detection is based on the following criteria. First, the sinusoid-to-residual energy ratio is below 5.0. Secondly, there are one or more peaks of ≥ -70dB[1] between 2.0 and 8.0kHz. Thirdly, the current-to-previous frame energy ratio is ≥ 1.5. When all three criteria are met, we switch to the Hann window of length 1024, and mark a region of 5 windows transient. Meanwhile, sine + residual decomposition is turned off. A typical transient region and the transient windows are depicted in Fig. 2.

2.3. Psychoacoustical peak detection

For each non-transient frame, its spectrum, computed using factor-8 zeropadding in time, is decomposed into a sinusoidal spectrum and a residual spectrum. The sinusoidal spectrum consists of up to 50 largest peaks above the masking threshold. The masking threshold is computed as the maximum of the human hearing threshold-

[1] Here, we normalize the maximum power to 0dB.

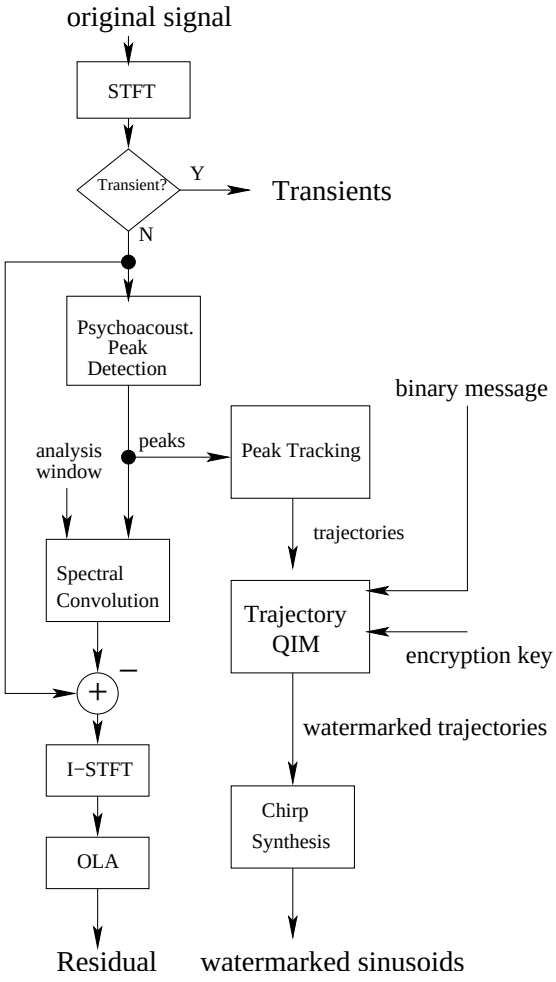

Figure 1: schematic diagram of watermark embedding

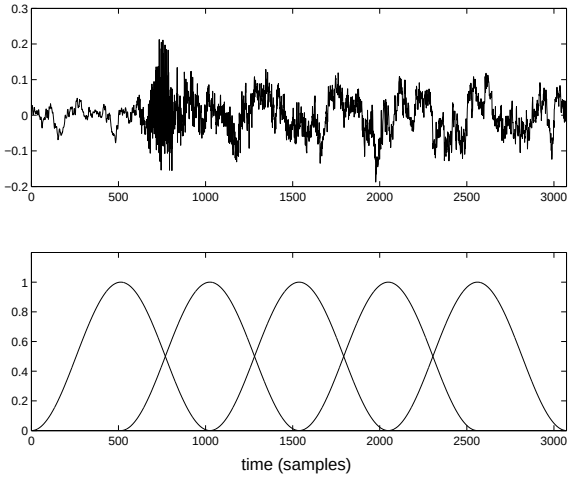

Figure 2: Transient detection and transient windowing

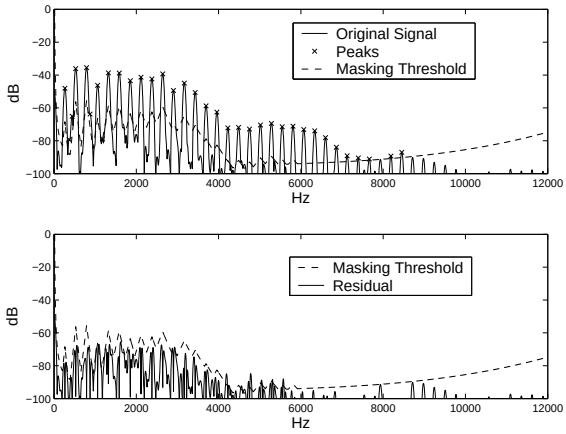

Figure 3: Psychoacoustical peak detection

in-quiet and spreading functions due to the presence of peaks. Each spreading function is described by two slopes and mimics the human hearing excitation pattern [6]. If a peak is identified tonal, its location (frequency), height (magnitude), and phase are estimated by parabolic interpolation in the *log* magnitude spectrum. These parameters are crucial for sinusoidal modeling, and the frequency will later be tracked and modified to carry information.

2.4. Computing the residual

For each tonal peak, a fitted mainlobe of the Blackman-Harris window transform is subtracted from the magnitude spetrum. This being done for all tonal peaks, the remaining is called the residual spectrum. The *residual part* is computed by overlap-adding (OLA) the inverse STFT (I-STFT) of all residual spectra. Fig. 3 demonstrates peak detection and residual computation.

2.5. Peak tracking

As described in section 2.3, peaks are estimated in frequency. Here, we shall describe how they are tracked in time to form *sinusoidal trajectories*. For every peak in the i^{th} frame, we try to connect it to the peak in the $(i+1)^{\text{th}}$ frame whose frequency is the closest. However, we do not allow jumps larger than 0.25 Bark. Neither are intersections and confluences of trajectories allowed. Because of these prohibitive rules, a trajectory can emerge and terminate as time proceeds.

2.6. Trajectory QIM: how information is embedded

To embed a watermark, frequencies of sinusoidal trajectories are slightly changed by quantization index modulation (QIM)[7]. As depicted in Fig. 4, two different sets of quantization grid points are associated with binary 0 and 1, called the quantization indice, respectively. For a frequency parameter to carry one bit of information, quantization points of index either 0 or 1 is used, depending

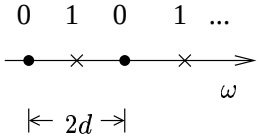

Figure 4: Quantization index modulation

on the watermark's binary value. To make the watermark transparent, the quantization size, defined as the distance d between neighboring points, is set to be just below human JND in frequency.

2.7. Chirp synthesis

We synthesize a watermarked sinusoidal part according to the trajectories that have QIM in frequency. The *sinusoidal synthesis* is based on linear chirps. Let (A_k, f_k, ϕ_0) be the amplitude envelope, the quantized frequency envelope, and the initial phase of a trajectory, where $k \in \{1, 2, ..., L\}$ is the frame number, and L is the trajectory length. Let M be the windowing hop-size. Then, the amplitude and the frequency are piece-wise linearly interpolated in time[2],

$$\begin{cases} f(t) = f_1, & 0 \le t < M \\ f(t) = (1-\lambda) f_k + \lambda f_{k+1}, & kM \le t < (k+1)M \\ f(t) = f_L, & LM \le t < (L+1)M \end{cases} \quad (1)$$

$$\begin{cases} A(t) = \lambda A_1, & 0 \le t < M \\ A(t) = (1-\lambda) A_k + \lambda A_{k+1}, & kM \le t < (k+1)M \\ A(t) = (1-\lambda) A_L, & LM \le t < (L+1)M \end{cases} \quad (2)$$

where $\lambda = t/M - k \in [0, 1)$ is a linear weighting factor. Also, the phase is the integral of the frequency.

$$\phi(t) = \phi_0 + 2\pi \sum_{\tau=1}^{t} f(\tau) \quad (3)$$

Accordingly, a *sinusoidal track* $s(t)$ is synthesized,

$$s(t) = A(t) \cos\big(\phi(t)\big),\ 0 \le t < (L+1)M \quad (4)$$

and the final sinusoidal part is the superposition of all sinusoidal tracks.

An example of trajectory watermarking and chirp synthesis is shown in Fig. 5. The original spectrum is also shown to compare with.

2.8. Final superposition

The sinusoidal, the residual, and the transient parts are superposed to form the final watermarked signal. Cares are taken to avoid artifacts when switching in and out of transient regions.

[2]For the simplicity of notations, we sloppily assume that $t = 0$ corresponds to the time each trajectory starts.

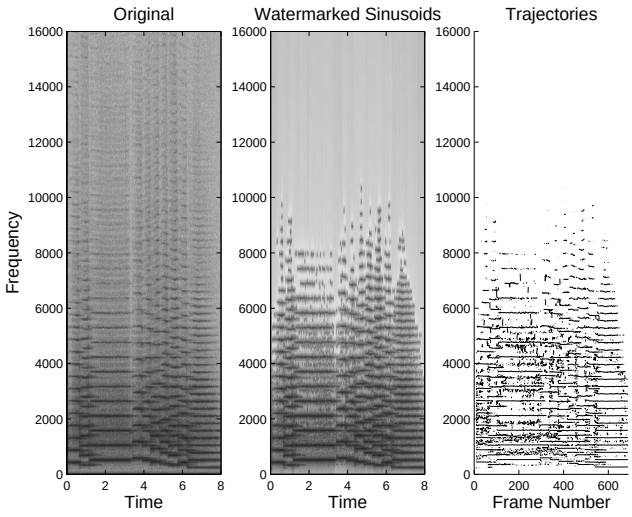

Figure 5: Trajectory watermarking

3. Watermark decoding

Because watermarks are embedded in the quantized frequencies of sinusoidal tracks, decoding the watermarks involves frequency estimation that is accurate enough to resolve adjacent quantization steps. From a received watermarked signal, frequencies of spectral peaks are first roughly estimated using the same methods as described in section 2.3. The frequency estimation is refined iteratively using a gradient descent method till the estimation converges [8]. Then, using quantization grid points corresponding to both quantization indice, at each decoding attempt, the hidden information is decoded as the binary quantization index to which a frequency is re-quantized. The decoding is successful if the frequency estimation error is below half of the quantization step size d.

4. Experiments

Here, we document attempts to hide information in 8.0 sec of a trumpet recording, whose spectrum is shown in Fig. 5. The targeted data hiding rate is one bit every 1024 samples, or 43 bps. The quantization step size d is set to be proportional to human JND in frequency, approximately fixed below 500Hz and linearly increasing above 500Hz [9]. For each frame in time, we embed a binary number W by using only one of two sets of QIM points to quantize all frequencies of concurrent trajectories. I.e., the quantization index is not modulated by all peaks independently in frequency, but is collaborately modulated by all of them in time. Before the watermarked signal is received by the watermark decoder, it goes through MP3 compression at 64 kbps. During the watermark decoding, let $\{(\hat{A}_i, \hat{f}_i) : i = 1, 2, 3, ...\}$ represent the estimated amplitudes and frequencies of spectral peaks in a frame. Let $\hat{b}_i \in \{0, 1\}$ be the index number to which $\hat{f}_i$ is requantized. Then, the final decoded binary value $\hat{W}$ for this

frame is given by the following weighted sum,

$$\hat{W} = \sum_i \alpha(\hat{A}_i) \cdot \hat{b}_i \quad (5)$$

where $\alpha(\cdot)$ is a strictly increasing function with the following property,

$$\sum_i \alpha(\hat{A}_i) = 1 \quad (6)$$

In Table 1, the decoding bit error rate (BER) averaged over $43 \times 8 = 344$ bits is shown at various d.

d (above 500Hz)	d (below 500Hz)	BER
10 cents	2.9Hz	10.5%
7 cents	2.0Hz	19.0%
5 cents	1.4Hz	26.9%
3 cents	0.87Hz	34.8%
2 cents	0.58Hz	39.8%

Table 1: Bit error rate versus quantization step size

5. Discussion and future directions

In Table 1, we can see the general trend that the BER increases when the quantization step size decreases. Nevertheless, it is debatable if the frequency quantization of 2 to 10 cents (of a semitone) introduces audible artifacts. Currently, we do not answer this directly but argue that, even though the difference between the original and the watermarked may be noticeable, the watermarked signal may not be objectionable if sine + residual + transient decomposition is implemented carefully. This is very likely the case if a subject has never listened to the original signal. However, further psychoacoustic tests have to be conducted before any conclusion can be drawn.

Also, the BER we obtained poses interesting questions on the usefulness of the watermark. For the 8 sec trumpet clip, current BER is small enough to serve as an evidence of the watermark's presence. Of course, the evidence is stronger as the quantization step size goes larger. However, the current BER is too high for the watermark to serve as a reliable communication channel at the targeted data hiding rate. We suspect that the major cause of BER is not MP3 compression, but the superposition of the residual part. In the future, we would like to improve on sinusoidal modeling. Hopefully, we can reduce the residual energy and the BER for a wider class of audio recordings.

6. Conclusions

Audio watermarks based on frequency quantization can be used to present the evidence of hidden information, even if the quantization size is below human JND. The embedding of such watermarks is possible because of a careful design of sinusoidal modeling, and the decoding of the watermarks requires very accurate frequency estimation. Our experiments have shown a certain level of robustness of the watermarks against MP3 compression. In the future, we believe that our sinusoidal model can be improved, and hope to develop the system for reliable data hiding at high rates.

References

[1] D. Kirovski and H. S. Malvar, "Spread-spectrum watermarking of audio signals," *IEEE Trans. Signal Processing*, vol. 51, no. 4, pp. 1020–1033, April 2003.

[2] M. van der Veen, F. Bruekers, J. Haitsma, T. Kalker, A. N. Lemma, and W. Oomen, "Robust, multifunctional and high quality audio watermarking technology," in *Audio Engineering Society 110th Convention (preprint)*, May 2001.

[3] S. N. Levine, "Audio representations for data compression and compressed domain processing," Ph.D. dissertation, Electrical Engineering Department, Stanford University (CCRMA), Dec 1998, available online.

[4] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan 1978.

[5] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, Nov. 1977.

[6] M. Bosi and R. E. Goldberg, *Introduction to digital audio coding and standards*. Kluwer Academic Publishers, 2003.

[7] B. Chen and G. W. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Information Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.

[8] Y.-W. Liu and J. O. Smith, "Watermarking sinusoidal audio representations by quantization index modulation in multiple frequencies," to be appeared in ICASSP 2004.

[9] T. D. Rossing, *The Science of Sound*. Addison-Wesley, 1990.