WATERMARKING PARAMETRIC REPRESENTATIONS FOR SYNTHETIC AUDIO

Yi-Wen Liu, Julius O. Smith III

Center for Computer Research in Music and Acoustics
Stanford University, Stanford, CA 94305, USA

ABSTRACT

This paper proposes to watermark parametric representations for synthetic audio. Our watermark system combines quantization index modulation at the encoder and maximum likelihood parameter estimation at the decoder. To guarantee error-free data hiding under expected types of attacks, knowledge of Fisher information and Cramér-Rao bounds is applied to the system design. Experiments show that, merely by quantizing the frequency of sinusoidal tones, one can achieve 50b/s of data hiding that is robust to perceptually shaped additive attacks such as an MP3 compression.

1. INTRODUCTION

Synthetic audio has found various applications in recent years. It is known that forward-looking audio coding standards, such as MPEG-4 structured audio [1], would allow a coder to explore concise algorithmic representations for sound synthesis. On the encoder side, this involves model selection based on audio content, and estimation of a set of parameters that best represent the signal. On the decoder side, a signal is parametrically synthesized. Interestingly, although "synthetic" literally means "not natural", there have been examples of synthetic speech [2], music [3], and random scenes [4] that sound very natural to human ears.

One may be interested in watermarking the parameters as well as the signal they generate, because parameters are expensive – it often takes a lot of computation power and human labor to obtain those that represent the signal well. Therefore, this paper concentrates on the design of watermarking schemes in the parameter space for audio syntheses.

Cox et al. [5] explicitly distinguishes between a watermark space and a signal space, and the former usually has a much lower dimensionality than the latter. Inspired by their work, this paper formulates an audio synthesis as a mapping from the watermark space to the signal space. Furthermore, if we model an attack in terms of its covariance in the signal space, its influences in the watermark space can be evaluated through an inverse mapping. It turns out that the distortion the attack causes in the watermark space has a Cramér-Rao bound (CRB), which implies that the quantization step sizes of an index modulation [6] watermarking scheme are similarly bounded below. Consequently, the achievable data hiding rate region is bounded above by the decoder's performance on parameter estimation. The closer it approaches the CRB, the higher the data hiding rate allowed.

The organization of this paper is as follows. Section 2 reviews Fisher information and the CRB, and gives a geometric interpretation for the case of parameter estimation in the presence of additive white Gaussian noise (AWGN). Section 3 proposes a parameter-space watermark system. Section 4 documents and discusses experiments on watermarking parameters for sinusoidal modeling synthesis subject to perceptually shaped attacks. Finally, future directions and conclusions are stated.

Symbols	Meanings
y_n, u_n	Signals in the time-domain. n is used as the time index consistently.
s_n^θ	Synthetic signal indexed by a set of parameters θ
$\hat{\theta}$	Estimate of θ
$\mathbf{y}$	Vector enumeration $(y_1, y_2, ..., y_N)$
$\mathbf{y} \sim f(\mathbf{y})$	Random variable with a distribution $f(\mathbf{y})$
$\mathcal{N}(\mathbf{m}, \Sigma)$	The normal distribution with mean $\mathbf{m}$ and covariance matrix Σ

Table 1. Notation

2. THEORY

2.1. Fisher Information and the CRB

Suppose that we are interested in estimating a single scalar parameter θ from the observation of a vector random variable $\mathbf{y} \sim f(\mathbf{y};\theta)$. Fisher Information $J(\theta)$ is defined as

$$J(\theta) = E\Big[\frac{\partial}{\partial\theta} \ln f(\mathbf{y};\theta)\Big]^2 \qquad (1)$$

and it can be interpreted as the amount of information about a parameter that can be extracted from an observation. This

can be explained by the Cramér-Rao inequality [7],

$$\mathrm{Var}(T(\mathbf{y}) - \theta) \geq \frac{1}{J(\theta)} \tag{2}$$

where T denotes an arbitrary unbiased parameter estimator. The inequality states that the expected mean square error of any unbiased estimator is bounded below by the CRB, which is defined as the reciprocal of Fisher information. However, there is no guarantee that a CRB-achieving estimator exists in general. Nevertheless, it can be shown [8] that if it exists, the estimator is a maximum likelihood (ML) estimator.

Also, the Cramér-Rao inequality can be generalized to a matrix version for the estimation of multiple parameters,

$$\Sigma \geq J^{-1}(\theta) \tag{3}$$

where Σ is the covariance matrix of the estimation error, and $J_{ij}(\theta) = E_{f(\mathbf{y};\theta)}\left[\frac{\partial \ln f}{\partial \theta_i}\frac{\partial \ln f}{\partial \theta_j}\right]$.

2.2. CRB for the estimation of audio synthesis parameters under Gaussian attacks

A *synthesis* is defined as a mapping from a *parameter space* $(\theta_1, \theta_2, ..., \theta_K)$ to a *signal space* $\mathbf{s}^\theta = (s_1, s_2, ..., s_N)$. If a synthesized signal $\mathbf{s}^\theta$ is subject to a Gaussian attack $\mathbf{u} \sim \mathcal{N}(0, \Sigma)$, we have

$$f(\mathbf{y};\theta) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \exp\{-\frac{1}{2}[\mathbf{y} - \mathbf{s}^\theta]^{\mathrm{T}} \Sigma^{-1} [\mathbf{y} - \mathbf{s}^\theta]\}$$

where $\mathbf{y} = \mathbf{s}^\theta + \mathbf{u}$ is the noisy observation. It can be shown [9] that the Fisher information is given by

$$J(\theta) = (\nabla \mathbf{s})^{\mathrm{T}} \Sigma^{-1} (\nabla \mathbf{s}) \tag{4}$$

where $(\nabla \mathbf{s})_i = \frac{\partial \mathbf{s}}{\partial \theta_i}$ may be called the sensitivity vectors.

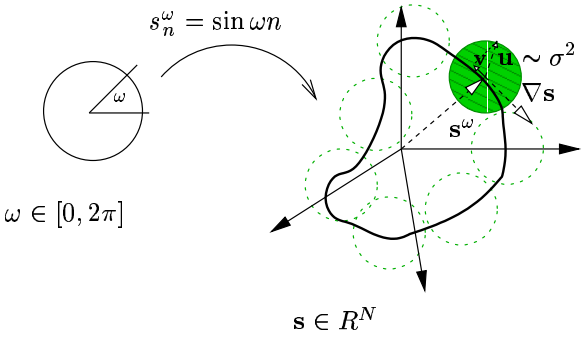

Fig. 1. Geometric interpretation of sinusoidal synthesis and AWGN attacks

To illustrate the concepts described above, Fig. 1 shows a geometric interpretation for the case when the synthesis is a single-tone sinusoidal model and the attack is AWGN $u_n \sim \mathcal{N}(0, \sigma^2)$. Here, the synthesis is a diffeomorphism from the unit circle $[0, 2\pi]$ to a curve $\mathbf{s}^\omega$ in the signal space. At each frequency ω, the attack is characterized by an N-dimensional ball of rms radius σ around $\mathbf{s}^\omega$. The ball blurs the resolution along $\nabla \mathbf{s}$, and the frequency error it causes may be approximated in the least squares sense. Define $\mathbf{v} = \delta\omega \frac{\partial \mathbf{s}}{\partial \omega}$ such that $\mathbf{v} \cdot (\mathbf{u} - \mathbf{v}) = 0$. Then, solving the pseudo-inverse problem, we have

$$E(\delta\omega)^2 = \frac{\sigma^2}{|\frac{\partial \mathbf{s}}{\partial \omega}|^2} \tag{5}$$

Interestingly, by comparing equations (4) and (5), it can be verified that,

$$E(\delta\omega)^2 = \frac{1}{J(\omega)}$$

in this special case. In other words, we can see Fig. 1 as a valid interpretation of Fisher information and CRB.

The next section describes a parameter-space audio watermarking system, in which the step sizes of the quantization codebooks are carefully chosen according to the CRB.

3. EMBEDDING AND DECODING ALGORITHMS

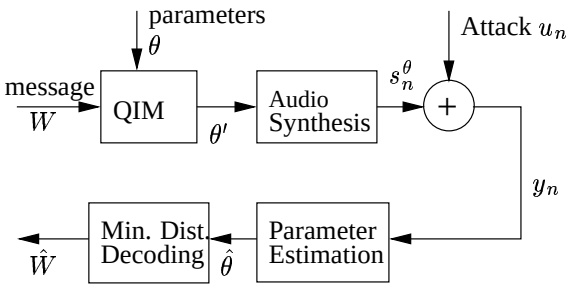

Fig. 2. System block diagram

Fig. 2 shows a generic parameter watermarking system for synthesized audio. On the encoder side, quantization index modulation (QIM) [6] is used to hide a message W in the parameter space. Watermarked parameters θ' are then fed into a synthesizer to generate an audio signal s_n^θ, which is subject to additive attack u_n. Upon reception of the distorted signal y_n, the decoder first estimates the parameters $\hat{\theta}$, and then uses minimum distance decoding to find a ML candidate message $\hat{W}$. An error occurs by definition if $\hat{W} \neq W$.

There are two constraints on the selection of quantization step sizes. Let d_i denote the minimum distance along the i^{th} parameter dimension between lattice points (See Fig. 3). First of all, the perceptual distortion $D_p(\mathbf{s}^{\theta_i} || \mathbf{s}^{\theta_i + d_i})$ should not be noticeable to human ears. Secondly, d_i should be large enough so that, under an attack, the decoder can still

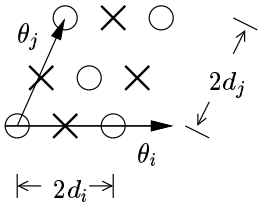

Fig. 3. Parameter-space quantization index modulation

distinguish a lattice point from its neighbors. Although this second constraint on d_i depends on the actual performance of the parameter estimation on the decoder side, we know from the inequality (2) that we need

$$d_i > (\alpha / J_{ii}(\theta)) \tag{6}$$

for robust data hiding, where $\alpha > 1$ controls the probability of error when we requantize to a d_i grid in the presence of noise.

Therefore, in practice, the encoder decides how much attack it would tolerate, and chooses quantization steps that satisfy the abovementioned constraints. Then, the decoder identifies the attack statistics and uses the best possible parameter estimator, which hopefully approaches the CRB.

4. EXPERIMENTS AND DISCUSSIONS

We present here experiments on single frequency QIM and ML frequency estimation to illustrate how the proposed watermarking system works. In particular, we intend the system to be robust against MP3 attacks. In this section, the synthesis is assumed to be a single tone sinusoidal model $s_n^\omega = \sin \omega n$.

4.1. CRB computation and frequency estimation under perceptually shaped attacks

For a sinusoid of a given frequency, we simulate the situation when the attack is additive Gaussian and spectrally shaped to the elementwise maximum of the following functions,

- A two-slope approximation [10] of the spreading function, which has the following form,

$$\begin{cases} S_{\text{left}} = 27 \ \text{dB/Bark}; \\ S_{\text{right}} = \begin{cases} -27, & \text{if } \Gamma < 40; \\ 27 + -0.37 * (\Gamma - 40) & \text{otherwise}, \end{cases} \end{cases}$$

  where S_{left} and S_{right} are the two slopes in units of dB/Bark, Γ is the magnitude in dB of the masker tone.

- A -80dB fixed SNR white noise floor.

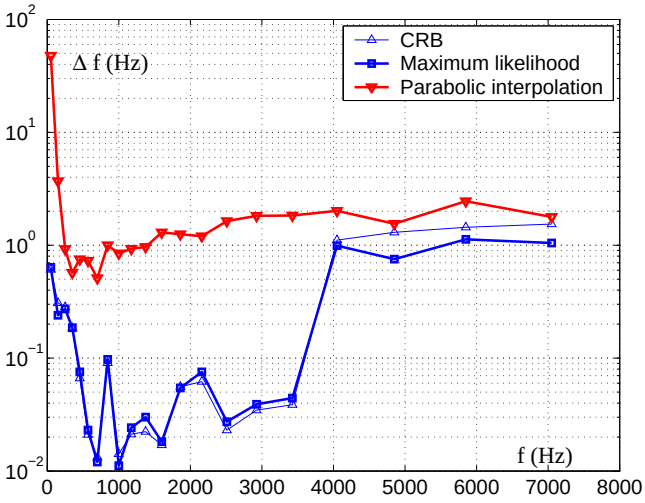

Fig. 4. Frequency estimation performances under psychoacoustically masked attacks

- A global hearing threshold function, which is above -80dB at low and high frequencies.

We implement a two-step ML frequency estimator and compare its performances with the CRB. The first step of the estimation is based on peak interpolation of the windowed Fourier transform. For each of 0.01s Hann windowed frame, a factor-8 zeropadding is used for calculating the Fourier transform. Then, the magnitude peak is estimated by parabolic interpolation. The interpolated peak frequency is regarded as a rough estimation, and the second step is a brute-force gradient descent search for ML frequency starting from the rough estimate. The step size of the gradient descent is set as $1/(2\sqrt{J(\omega)})$. The sampling rate is 16kHz.

Fig. 4 shows the simulation results. The horizontal axis shows the frequency of the sinusoid, and the vertical axis shows the root mean square frequency estimation error. For each run, the frequency is set to be the center frequency of a bark band. The plot is obtained from an average over 20 runs. At low and middle range frequencies, the performance of the coherent ML estimator significantly outperforms the non-coherent parabolic magnitude interpolation. However, when the masker sinusoid has a high frequency, the improvement is marginal because the attack is more spectrally flat, based on the two-slope approximation of the spreading function.

Finally, we are aware that, in this plot, the estimation error at high frequencies lies below the CRB. It could possibly be due to that the estimator is biased, but this is not clear to us yet.

4.2. Frequency QIM against MP3 attacks

Nevertheless, the frequency estimator looks accurate and robust enough against perceptually shaped attacks in simulations. In a more realistic test, we implement frequency QIM with two equally spaced lattices. The quantization step size d is set as 1 or 2 Hz. In our informal listening tests, frequency quantization of such step sizes does not cause significant perceptual distortion to our test signal, which is a simple tune widely used as cell phone ringers. A commercially available MP3 attack is inserted between the encoder and the decoder. The MP3 attack compresses the signal to 9-12 kbps. The two-step frequency estimator is used at the decoder. The experiment results are summarized in Table 2, where the first column shows the quantization step size d, the second column shows the attempted data hiding rate R, and the third column shows the bit error rate P_e as results. All the error rates are obtained over an average of 2200 attempts.

d	R	P_e
2 Hz	50 b/s	0.14%
1 Hz	50 b/s	3.73%
2 Hz	100 b/s	5.59%

Table 2. Data hiding rates of frequency QIM under MP3 attacks

5. FUTURE DIRECTIONS

So far, we have proposed a generic audio parameter-space watermarking system, but only experimented with single parameter syntheses using simple sinusoidal models. In the future, we would like to extend the current work to the watermarking of multiple parameters that are used for various types of syntheses. We expect challenges on both the encoding and the decoding sides due to the much richer geometry of the synthesis mappings. In particular, we are interested to see if QIM on a curved parameter space is still straightforward. Also, it may also be nontrivial to design a gradient descent ML estimator on a curved manifold in the signal space.

6. CONCLUSION AND SUMMARY

This paper focuses on watermarking parametric representations for synthetic audio signals. The proposed system uses parameter-space quantization index modulation at the encoder and maximum-likelihood estimation at the decoder. The quantization step size has to be large enough so that the parameter estimator can distinguish between lattice points blurred by attacks; but not so large that there is distortion noticeable to human ears. The system is tested on the case of watermarking the frequency parameter of a sinusoidal synthesis subject to MP3 attacks, and the results show that 50b/s of reliable and imperceptible data hiding can be achieved.

7. REFERENCES

[1] Barry L. Vercoe, William G. Gardner, and Eric D. Scheirer, "Structured audio: Creation, transmission, and rendering of parametric sound representations," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 922–940, May 1998.

[2] J. Schroeter, J. Ostermann, H.P. Graf, M. Beutnagel, E. Cosatto, A. Syrdal, A. Conkie, and Y. Stylianon, "Multimodal speech synthesis," in *Proceedings of IEEE International Conference on Multimedia and Expo*, New York, Jul. 2000, pp. 571–574, IEEE Press.

[3] Scott Nathan Levine, *Audio Representations for Data Compression and Compressed Domain Processing*, Ph.D. thesis, Electrical Engineering Department, Stanford University (CCRMA), Dec 1998, available online at http://www-ccrma.stanford.edu/s̃cottl/papers.html.

[4] Perry Cook, "Modeling bill's gait: Analysis and parametric synthesis of walking sounds," in *Proceedings of the AES 22nd International Conference on Virtual, Synthetic, and Entertainment Audio*, Jun. 2002, pp. 73–78.

[5] Ingemar J. Cox, Matthew L. Miller, and Andrew L. McKellips, "Watermarking as communications with side information," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1127–1141, Jul. 1999.

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

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York, 1991.

[8] Harry L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*, Wiley-Interscience, New York, 2001.

[9] Louis L. Scharf and L.T. McWhorter, "Geometry of the Cramér-Rao bound," in *Proceedings of IEEE 6th signal processing workshop on statistical signal and array processing*, 1992, pp. 5–8.

[10] Marina Bosi, "Perceptual audio coding," *IEEE Signal Processing Magazine*, vol. 14, no. 5, pp. 43–49, Sep. 1997.