A COHERENT MULTIPLE FREQUENCY ESTIMATOR AND ITS APPLICATIONS IN SYNTHESIS-ANALYSIS AUDIO WATERMARKING

Yi-Wen Liu and Julius O. Smith

Stanford University
Center for Computer Research in Music and Acoustics
660 Lomita Dr., Stanford, CA 94305, USA
{jacobliu, jos}@ccrma.stanford.edu

ABSTRACT

This paper extends our previous work on single parameter watermarking to the multiple parameter cases. Watermarks are embedded in multiple sinusoidal tracks via frequency quantization index modulation (F-QIM), and extracted by a coherent multiple frequency estimator using Newton’s method. Experiments are conducted subject to MP3 attacks and, in an 8-track sinusoidal synthesis-analysis setup, above 400 bits/sec of data hiding is consistently achieved with error correction. Furthermore, the frequency estimator’s performance is compared against the fundamental limit of the inverse Fisher information.

1. INTRODUCTION

Synthesized multimedia objects are emerging everywhere. One can talk on the phone to a virtual representative, drink soda of synthesized taste, or even fall in love with computer graphic characters. It becomes urgent to protect such objects as intellectual properties, for synthesizing them often involves a lot of computation power and human labor. In recent years, works have been done to watermark graphical objects (e.g. [1, 2]) directly in the parameter spaces. Somewhat parallel to these works, in [3], we proposed a data hiding scheme to watermark parametric representations for synthetic audio.

As shown in Fig. 1, the scheme combines a quantization index modulation (QIM)[4] encoder and a maximum-likelihood (ML) decoder. While the scheme has been validated by successfully hiding data at 50 bit/s in pure single sinusoids under MP3 attacks, in this paper, we extend to the multiple sinusoid case and improve the data hiding rates by an order of magnitude.

The organization of this paper is as follows. In Section 2, a mathematical definition of parametric synthesis is presented. Also, Fisher information and the Cramér-Rao inequality are reviewed to explain the limit of parameter estimation in general and parameter space watermark decoding in particular. In Section 3, a coherent multiple frequency estimation algorithm is proposed. In Section 4, the performance of the algorithm is evaluated by comparing frequency estimation error covariance against the inverse Fisher information matrix. Finally, experiments are documented in Section 5.

2. MATHEMATICAL FORMULATION

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Meanings</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_n, u_n, v_n$</td>
<td>Signals and noise in the time-domain. $n$ is the time index.</td>
</tr>
<tr>
<td>$s_n^\theta$</td>
<td>Synthetic signal indexed by a vector parameter $\theta = (\theta_1, \theta_2, ..., \theta_K)^T$. $\theta$ may be dropped whenever there is no confusion.</td>
</tr>
<tr>
<td>$\mathbf{y}$</td>
<td>Vector enumeration $(y_{-N}, ..., y_N)^T$</td>
</tr>
<tr>
<td>$\mathbf{y} \sim f(\mathbf{y}; \theta)$</td>
<td>Vector random variable $\mathbf{y}$ with a distribution $f$ parametrized by $\theta$.</td>
</tr>
<tr>
<td>$\hat{\theta}$</td>
<td>Estimate of $\theta$. $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K)$</td>
</tr>
</tbody>
</table>

Table 1: Notations used throughout this paper

Table 1 summarizes the mathematical notations. An audio synthesis is defined as a function that takes a vector parameter $\theta \in \mathbb{R}^K$ as the input and maps it to a signal...
Δf (Hz)
Figure 2: Performance of simultaneous tracking of two well-separated sinusoids. (a) A₁ = A₂ = σ, f₁ = 440Hz, f₂ = 240Hz. (b) A₁ = A₂ = 10σ, f₁ = 440Hz, f₂ = 240Hz.

s° ∈ C^{2N+1}. Since our watermarking scheme involves parameter estimation, it is interesting to study if there is any fundamental limit to the estimation accuracy.

Let y be an observation of the synthesized signal s after an additive attack u, and let u be probabilistic such that

\[ y = s + u \sim f(y; \theta) \]  

The Fisher information matrix J is defined as follows,

\[ J_{ij}(\theta) = E_{f(y; \theta)} \left[ \frac{\partial \ln f}{\partial \theta_i} \frac{\partial \ln f}{\partial \theta_j} \right] \]  

It can be shown [5] that the Fisher information matrix is the Riemannian metric of the K-dimensional manifold of parametrized distributions f(y; θ).

\[ D(\theta|\theta + d\theta) = \frac{1}{2} d\theta^T J(\theta) d\theta \]  

where

\[ D(\theta||\theta') = \int f(y; \theta) \log \frac{f(y; \theta)}{f(y; \theta')} dy \]  

is the Kullback-Leibler divergence, which defines the distance between any two points in the manifold.

Because \( D(\cdot||\cdot) \) can be interpreted as the discrepancy between probability distributions, when distributions are continuously parametrized, the Fisher information matrix, which is the curvature tensor associated with \( D(\cdot||\cdot) \), can qualitatively be interpreted as how much information \( \theta \) reveals about itself along every direction in the K-dimensional parameter space. Quantitatively, this argument is justified by the Cramér-Rao matrix inequality [6],

\[ \Sigma^θ \geq J^{-1}(\theta) \]  

where \( \Sigma^θ = \text{Cov}(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T \) is the covariance matrix of the estimation error by any unbiased estimator \( T(y) = \hat{\theta} \).

In the case of multiple frequency estimation under white Gaussian attacks, the parameter vector \( \theta = (\omega_1, \omega_2, \ldots, \omega_K) \) consists of all the unknown frequencies. Let the synthesized signal be

\[ s_{n+1}^{ω_1, ω_2, \ldots, ω_K} = \sum_{k=1}^{K} A_k \exp[j(ω_k n + φ_k)] \]

and let \( u_n \) be in C and have i.i.d. Gaussian real and imaginary parts with \( \mathcal{N}(0, σ^2) \). Let \( n \in [-N, N] \) be the time frame for observation. Then we have

\[ f(y; \theta) = \frac{1}{\sqrt{2πσ^2}} \exp\left(-\frac{1}{2σ^2} |y - s^θ|^2 \right) \]  

Now, using Eq.(2), the Fisher information matrix can be
In particular, with \( i = j \), the diagonal terms have the following simple close form expression,

\[
J_{ii} = \frac{A_i^2}{\sigma^2} \cdot \frac{N(N+1)(2N+1)}{3}
\]  

(9)

In the special case if all the unknown frequencies are well separated, we have

\[
|\omega_i - \omega_j| \cdot N \gg 2\pi, \forall i \neq j
\]

(10)

and asymptotically,

\[
J_{ij} = O(N^2) \ll J_{ii}, \forall i \neq j
\]

(11)

Then, inequality (5) loosely says that \( J^{-1} \) consists of the dominating diagonal elements \( 1/J_{ii}, i = 1, 2, ..., K \), which each is a lower bound to \( E((\omega_i - \omega_j)^2) \), the individual frequency estimation variance. Moreover, the bound can be achieved using a coherent ML estimator [7].

However, if any two of the unknown frequencies are close to each other, the corresponding off-diagonal elements \( J_{ij} \) can not be ignored. We shall examine this in Section 4 and study its implication to the design of sinusoidal synthesis-analysis watermarking systems.

3. METHODS

3.1. Watermark encoding

For a set of sinusoidal synthesis parameters, we attempt to hide one bit per frame per track in the frequencies. For each frequency as a parameter, two interleaving scalar codebooks are used for F-QIM. The spacing between codebooks, or the quantization step size, should be large enough so that watermarks survive certain expected types of attack, and small enough so that the watermarks are not audible. We empirically choose the spacing as a fixed 2Hz below 500Hz, and 10 cents of a semitone above 500Hz. Informal listening tests confirm that the artifacts introduced by the watermarks are acceptable, if not imperceptible to the authors.

3.2. Watermark detection

The frequency estimation consists of two stages. In the coarse stage, the short-time spectrum of \( y \) is calculated using the Gaussian window [8] with \( \alpha = 2 \). Then, we use parabolic interpolation of the log magnitude spectrum to locate \( K \) peaks simultaneously. If this fails, we enlarge the scope by switching to a longer window until \( K \) peaks are found. The frame hopping step is kept unchanged to avoid synchronization difficulties. The result of coarse estimation initializes the fine stage specified as the following,

Step 0: Given an observation \( y \), let \( \xi = \xi(\omega_1, \omega_2, ..., \omega_K; y) \) be the least square error of \( y \) with respect to \( \text{span} \{\cos(\omega_j n), \sin(\omega_j n) : j = 1, 2, ..., K\} \).

Step 1: Initialize with \( (\omega_1, \omega_2, ..., \omega_K) \) obtained from the coarse stage.

Step 2: Set \( i = 1 \).

Step 3: Measure the 1\textsuperscript{st} and 2\textsuperscript{nd} partial derivatives of \( \xi \) with respect to \( \omega_i \).

Step 4: Fix the frequencies \( \omega_j, j \neq i \), and update \( \omega_i \) by Newton’s method [9].

\[
\omega_i \leftarrow \omega_i - \left( \frac{\xi^+ - \xi^-}{\xi^+ - 2\xi^+ + \xi^-} \right) \cdot \Delta \omega
\]

(12)

where \( \Delta \omega \) is a small perturbation for derivative measurements.

Step 5: Next \( i \leftarrow i + 1 \mod K \). Go to Step 3.

After frequency estimation, minimum distance watermark decoding is guaranteed successful if the estimation error is less than a half of the quantization step size. In the next section, we shall study when this does and does not occur.

4. PERFORMANCE ANALYSIS

In this section, the performance of the two-stage frequency estimator is analyzed. We shall examine if the fine-stage converges and improves the coarse estimation. Also, provided that it converges, it is interesting to see if the frequency estimation error approaches the Cramer-Rao lower bound (CRB).

Simulation results are shown in Fig. 2. Various sets of SNR and frequency spacing are tested with a fixed sampling rate of 16kHz and frame rate of 62.5 per second. I.e., the frame length is 16ms and \( N = 128 \). In the plots, the ellipses in solid line represent the nominal bias and variance of the coarse stage of frequency estimation over 100 runs. The dash-dot ellipses represent the fine estimation after 5 iterations. The ellipses marked with circles represent the inverse of Fisher information. The actual estimated frequency pairs are each marked with \( x \). From Fig. 2(a) and (b), it is clear that the fine stage helps to reduce both the bias and the variance of error when frequencies are well separated. Fig. 2(b) shows that the nominal estimation error is within 1Hz subject to attacks around 20dB of SNR. Also, the estimator is quite efficient in terms of the speed approaching CRB.

However, when frequencies can not be well resolved within a frame length, as shown in Fig. 3, Newton’s method tends to diverge even if the coarse stage gives a good initial frequency estimation due to window switching. It is not...
the estimator is capable of detecting F-QIM watermarks at a meaningful rate to converge. Nevertheless, we empirically conclude that the previously described methods are tested for watermarking audio subject to attacks by a commercially available MP3 codec. We sinusoidally synthesized the first two measures of a 4-part orchestration of Air in D from Suite No. 3 composed by J.S. Bach. Each of the 4 instruments spectrally consists of one fundamental and one harmonic. The two partials have comparable amplitudes. The synthesis frame rate is 62.5 per second, and the synthesized audio has a sampling rate of 16kHz. The MP3 attack compresses the audio to about 18 kbps.

Fig. 4 shows a typical frequency estimation result. As expected, the estimation is quite successful when the 8 frequency trajectories do not collide with one another. The error is most prominent between the 280th and the 340th frame, when the 3rd and the 4th tracks from the top (Tenor’s 2nd harmonic and Soprano’s fundamental) actually have the same frequency.

Experiments show that the average watermark decoding bit error rate $P_{e,i}$ of track $i$ over 2560 frames ranges from $0.32\%$ (6th track) to $4.76\%$ (3rd track). It is promising that error correction codes can be used to achieve a data hiding rate $R = 62.5 \times \sum_{i=1}^{8} 1 - H(P_{e,i})$ bits per second, where $H(p)$ is the entropy of Bernoulli process with probability $p$, and $1 - H(p)$ is the capacity of the corresponding binary symmetric channel. The average $R$ we obtained over the 2560 frames is 437 bits/sec.

6. CONCLUSION

We have presented a two-stage multiple frequency estimator for parameter space watermarking. The first stage gives a rough frequency estimation that is refined by the coherent detection stage using Newton’s method. The frequency estimator successfully converges and approaches CRB when the target frequencies are well separated. The estimator is used in decoding F-QIM watermarks and helps achieve above 400 bps of data hiding in an 8-track sinusoidal synthesis.

7. REFERENCES


