WATERMARKING SINUSOIDAL AUDIO REPRESENTATIONS BY QUANTIZATION INDEX MODULATION IN MULTIPLE FREQUENCIES

Yi-Wen Liu, Julius O. Smith

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

ABSTRACT

The focus of this paper is to hide information in sinusoidal audio representations. Watermarks are embedded by introducing quantization index modulation (QIM) in frequencies of sinusoids. The frequency shifts due to QIM are controlled to be less than one-thirtieth Bark, which requires high accuracy in frequency estimation for watermark decoding. Therefore, this paper presents a hybrid multiple frequency estimator which consists of spectral interpolation followed by iterative square error minimization. The frequency estimator is tested in watermarking both computer generated sounds and actual recordings. The watermarks are intended to survive MP3 compression. Selected experiments are documented and discussed.

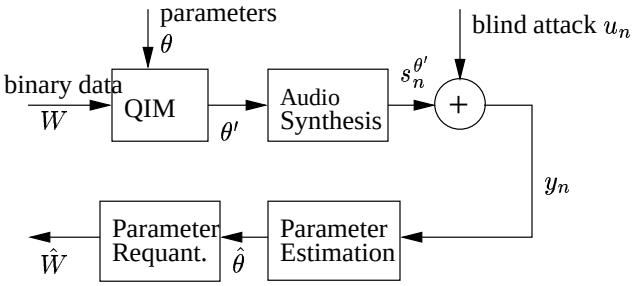

Fig. 1. Watermarking parametric representations for synthetic audio

1. INTRODUCTION

In [1], we proposed an information hiding framework to watermark parametric representations for synthetic audio. QIM, as explained in details in Sec. 2, is introduced to the parametric representation instead of the audio signal itself. Quantized parameters, hence carrying information, are used to synthesize a watermarked version of the signal. The watermarked signal can go through blind operations that introduce no perceptible artifacts, and the watermarks are designed to remain decodable after such operations. Under this framework, it was demonstrated in [1] how frequencies of sinusoids as parameters can be used for watermarking. In an example of watermarking single tone sinusoids, a data hiding rate of 50 bps was achieved subject to MP3 compression with relatively low bit error rate (BER).

However, during watermark decoding, the result above was obtained by exhaustive maximum-likelihood frequency estimation, which is inefficient when the parameter space is multi-dimensional. To extend from single-tone to multi-tone watermarking, this paper proposes a fast iterative method for multiple frequency estimation. The method uses parabolic interpolation of the logarithmic magnitude spectrum to obtain an initial estimate of peak frequencies, and the estimation is refined by a few iterations using Newton's method. Although Newton's method is well known to be not always stable, it is shown in Sec. 3 empirically that, if frequencies are well resolved to begin with, parabolic spectral interpolation gives a good initial estimation for subsequent iterations to converge and reduce the estimation error. The performance of the frequency estimator is compared against the theoretic limit of the Cramér-Rao lower bound, which is derived in the appendix.

Finally, the frequency estimator is tested in watermarking sinusoidal representations for a computer generated sound and for an actual recording. In the latter case, the sinusoidal representation does not exist a priori and has to be transcribed by sine + noise + transient decomposition [2]. Not surprisingly, much better data hiding rate and BER are obtained in the case of computer generated sound. Sec. 4 documents the test results, and Sec. 5 describes several challenges to overcome before frequency QIM can be applied for watermarking general audio.

2. SYSTEM OVERVIEW AND DEFINITION OF TERMINOLOGIES

Fig. 1 shows a generic block diagram of watermarking parametric representations for synthetic signals. Throughout this paper, the vector parameter θ is assumed to be the frequencies $(\omega_1, \omega_2, ..., \omega_K)$ of multiple sinusoids.

2.1. Watermark encoding based on QIM

For each sinusoid, we attempt to hide one bit per frame length by introducing QIM to its frequency. In other words, as shown in Fig. 2, one of two interleaving sets of frequency quantization points are used depending on the binary value the sinusoid is supposed to carry. The quantization step size, defined here as the spacing d between codebooks, should be large enough so that watermarks survive expected types of signal processing, and small enough so that the frequency shifts are not objectionable. The spacing is linear below 500Hz, and log scaled above 500Hz, which is roughly proportional to the spacing of critical bandwidths. In the experiments described later, the spacing is chosen to be between $\frac{1}{60}$ and $\frac{1}{30}$ Bark, which is comparable to the just noticeable difference (JND) in human pitch perception [3].

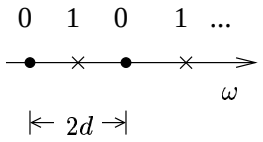

Fig. 2. Quantization index modulation

2.2. Audio synthesis

Given quantized frequencies, the sinusoids are synthesized in a phase-continuous manner to avoid artifacts at frame rate.

2.3. Blind signal processing

Currently, we only consider making watermarks robust to common signal processing procedures that introduce no perceptible distortions, particularly the MP3 compression. Also, within the scope of this paper, synchronization attacks are not considered.

2.4. Frequency estimation

The frequency estimation consists of two stages. In the coarse stage, the short-time spectrum of observed signal $\mathbf{y} = (y_{-N}, y_{-N+1}, .., y_N)$ is calculated using the Gaussian window [4] with $\alpha = 2$. Then, assuming the number of sinusoids is known to be K, the log magnitude spectrum is parabolically interpolated to locate K peaks simultaneously. If this fails, the window length keeps increasing until K peaks are found. The frame rate remains the same as in watermark encoding. The interpolated peak frequencies initializes the fine stage, which is supposed to iteratively minimize the cost function $\xi = \xi(\omega_1, \omega_2, ..., \omega_K; \mathbf{y})$ defined as the least square linear approximation error of $\mathbf{y}$ with respect to the subspace spanned by the sinusoids of frequencies $(\omega_1, \omega_2, ..., \omega_K)$.

$$\xi(\omega_1, \omega_2, ..., \omega_K; \mathbf{y}) = \left| \left(\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T - \mathbf{I} \right) \mathbf{y} \right|^2, \quad (1)$$

where the matrix $\mathbf{A}$ consists of $2K$ real-valued column vectors corresponding to sin and cos of frequencies $(\omega_1, \omega_2, ..., \omega_K)$. Then, the iteration procedure is specified as the following,

`Step 1:` Initialize with the frequencies obtained from the coarse stage. Set $i = 1$.

`Step 2:` Measure the 1^{st} and 2^{nd} partial derivatives of ξ with respect to ω_i by introducing a small perturbation $\Delta\omega$.

`Step 3:` Fix all other frequencies $\omega_j, j \neq i$, and update ω_i by Newton's method [5].

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

`Step 4:` Increase i by 1. Go back to `Step 2`.

An iteration is defined as an entire round in which all frequencies are updated once.

2.5. Watermark decoding

The estimated frequencies are re-quantized to decode the watermark binary values. Decoding is successful if the frequency estimation error is within $\pm\frac{d}{2}$.

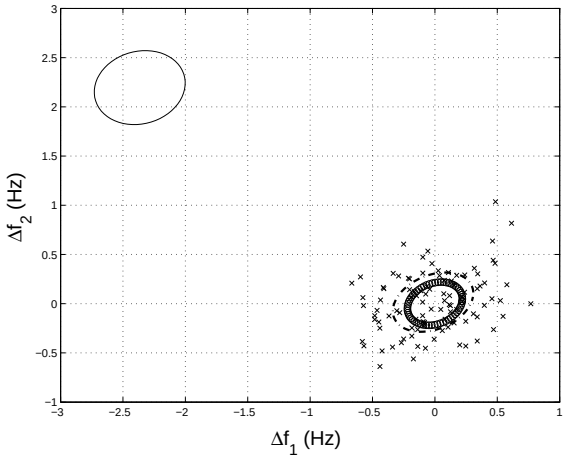

Fig. 3. Simultaneous frequency estimation under AWGN. $f_1 = 440$Hz, $f_2 = 240$Hz, and SNR = 20dB.

3. FREQUENCY ESTIMATION PERFORMANCES

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.

Various SNR and frequency spacing are set to test the frequency estimator with the fixed sampling rate of 16kHz and frame length of 16ms. Fig. 3 shows the result of one typical setting. The signal consists of two sinusoids that are well resolved within the frame length. The two sinusoids have the same amplitude. Additive white Gaussian noise (AWGN) is added to the signal before frequency estimation begins. In the plot, the solid-lined ellipse shows the nominal bias and variance after the coarse stage of frequency estimation, averaged over 100 attempts. The dash-dot ellipse corresponds to the fine estimation after 5 iterations. The ellipse marked with circles shows the Cramér-Rao bound (CRB) [6]. The 100 actual estimated frequency pairs are each marked with $\times$.

From Fig. 3, it is clear that the fine stage helps to reduce both the bias and the variance. The nominal estimation error is within 1Hz at 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, Newton's method tends to diverge even if the coarse stage gives a good initial frequency estimation. In this case, a more sophisticated frequency estimator is necessary and left as a future research direction.

4. EXPERIMENTS

The previously described system is tested in watermarking subject to compression and decompression by an MP3 codec.

4.1. Watermarking a computer-generated signal

A sinusoidal representation is manually transcribed for 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 a fundamental and an overtone. The two partials have comparable amplitudes. Below and above 500Hz, the QIM step

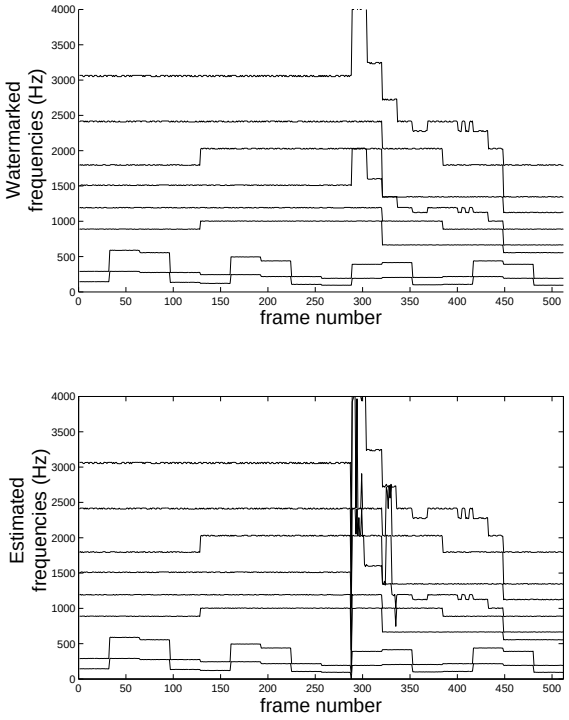

Fig. 4. Frequency trajectories of Bach's Air in D and frequency estimation after MP3. The top plot shows the watermarked frequency trajectories. The bottom shows the result of frequency estimation.

size is 2Hz and 10 cents (of a semitone), respectively. The 8 sinusoids are quantized independently. The synthesis frame rate, which is the data hiding rate per sinusoid, is 62.5/s, and the synthesized audio is sampled at 16kHz. The MP3 codec compresses the signal 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 frames, when the 3rd and the 4th trajectories from the top actually have the same frequency.

Experiments show that the watermark decoding bit error rate $P_{e,i}$ of trajectory i, averaged over 2560 bits, ranges from $P_{e,6} = 0.32\%$ to $P_{e,3} = 4.76\%$.

4.2. Watermarking a recorded signal

In Fig. 5, the spectrogram of an actual trumpet recording is shown on the left. On the right, the spectrogram of its sinusoidally modeled components, with frequency QIM, is shown. The details of sinusoidal modeling are beyond the scope of this paper. It suffices to summarize here that a sine + noise + transient decomposer constantly looks for the best representation of the signal as a superposition of quasi-stationary sinusoids, and the sinusoids can emerge, disappear, smoothly change frequencies and amplitudes, and be interrupted by regions of rapid transient in time. More interested readers are directed to [2].

In this example, the step size of frequency QIM is 1.5Hz below and 5 cents above 500Hz. The residual of sinusoidal modeling is

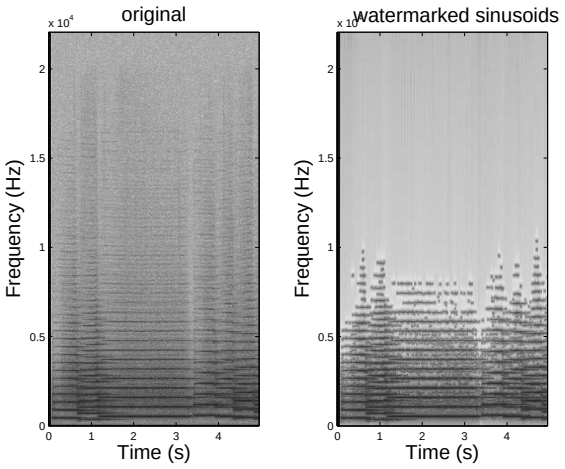

Fig. 5. A spectrogram and its frequency QIM version

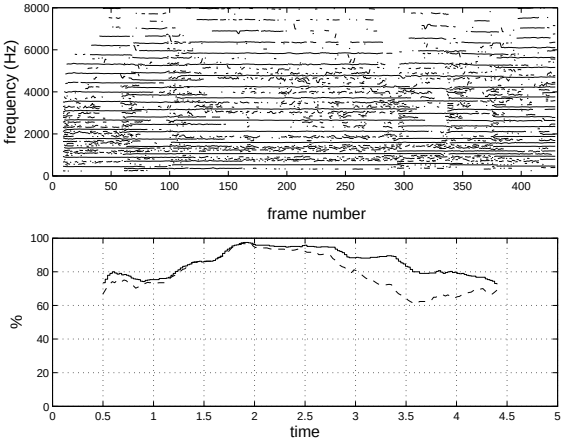

Fig. 6. Frequency trajectories and watermark decoding accuracy

added back to the watermarked sinusoids to form the watermarked version of the original signal. No transient regions are detected.

43 bps of data hiding rate is attempted by joint frequency QIM of all sinusoids above the masking threshold computed using a simple two-slope spreading function [7]. Here, joint frequency QIM means that, at each frame, all sinusoids are quantized using the same codebook. Again, the watermarked signal goes through the MP3 codec. On the decoding side, the estimated watermark binary value is an energy weighted sum of binary values determined by individual sinusoids.

Fig. 6 shows the success rate of watermark decoding. On the top, frequency trajectories are shown as a reference. On the bottom, $1 - \mathrm{BER}$ is plotted by smoothing over a period of 1.0s and averaging over an ensemble of 10 runs. The success rate of the first and the last 0.5 second is not shown due to time-smoothing. The solid line shows the decoding accuracy with Newton's iterations for frequency estimation, and the dashed line shows the accuracy without iterations. It is clear that Newton's iteration helps to reduce BER. However, the BER is much higher than in the case of watermarking computer generated sound, and strongly depends on signal characteristics as time proceeds.

5. DIAGNOSES AND FUTURE DIRECTIONS

In the future, we would like to study robust frequency estimation for under-resolved cases. Also, for watermarking based on frequency QIM to be applied to more general audio, our current sinusoidal modeling technique is not sophisticated enough to handle chirps well. After all, the very idea of parabolic interpolation and iterative phase locking is based on the assumption of stationarity. As a result, the residual of sinusoidal modeling is not negligible when frequencies vary rapidly, and comes in as extra noise when it is added back to the signal, which severely degrades watermark decoding performances. Therefore, we would like to pursue a higher level of signal modeling. Finally, since the human frequency JND varies from person to person, it is debatable if a difference of $1/60$ Bark can be heard. We are interested in a thorough psychoacoustic test to evaluate if the frequency shifts introduced by the proposed watermarking system are noticeable and objectionable.

6. CONCLUSION

The hybrid multiple frequency estimator is shown to converge and approach CRB when the frequencies to estimate are well resolved within the time frame. The estimator is applied to the decoding of watermarks that are embedded in frequencies of sinusoidal audio representations. The embedding is based on QIM, and the decoding is sucessful if the frequency estimation error is below half of the quantization step size. Experiments show that, by introducing debatably noticeable frequency modulation, audio data hiding can be achieved subject to MP3 compression. The actual data hiding rate depends on the frequency spacing of sinusoids, the number of sinusoids, and, in the case of watermarking an actual recording, how well the signal can be sinusoidally modeled.

7. REFERENCES

[1] Yi-Wen Liu and Julius O. Smith, "Watermarking parametric representations for synthetic audio," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, Hong Kong*, Apr. 2003.

[2] 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/ scottl/papers.html.

[3] Thomas D. Rossing, *The Science of Sound*, Addison-Wesley, 1990.

[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] Bernard Widrow and Samuel D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, New Jersey, 1985.

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

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

A. DERIVATION OF CRB FOR MULTIPLE FREQUENCY ESTIMATION UNDER AWGN

Define an audio synthesis as a function that maps a vector parameter $\theta \in \mathbb{R}^K$ to a signal $\mathbf{s}^\theta \in \mathbb{C}^{2N+1}$. Let $\mathbf{y}$ be an observation of $\mathbf{s}^\theta$ subject to an additive noise $\mathbf{u}$, and let $\mathbf{u}$ be probabilistic such that

$$\mathbf{y} = \mathbf{s}^\theta + \mathbf{u} \sim f(\mathbf{y}; \theta) \tag{3}$$

The Fisher information matrix J is defined as follows,

$$J_{ij}(\theta) = E_{f(\mathbf{y};\theta)}\Big[\frac{\partial \ln f}{\partial \theta_i}\frac{\partial \ln f}{\partial \theta_j}\Big] \tag{4}$$

The Cramér-Rao matrix inequality [6] sets the lower bound of error variance for non-biased estimators,

$$\Sigma^\theta \geq J^{-1}(\theta) \tag{5}$$

where $\Sigma^\theta = \mathrm{E}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T]$ is the covariance matrix of the estimation error by any unbiased estimator $T(\mathbf{y}) = \hat{\theta}$.

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

$$s_n^{\omega_1,\omega_2,...\omega_K} = \sum_{k=1}^{K} A_k \exp[j(\omega_k n + \phi_k)] \tag{6}$$

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

$$f(\mathbf{y}; \theta) = \frac{1}{\sqrt{(2\pi\sigma^2)^{2(2N+1)}}} \exp(-\frac{1}{2\sigma^2}|\mathbf{y} - \mathbf{s}^\theta|^2) \tag{7}$$

Now, using Eq.(4), the Fisher information matrix can be derived,

$$J_{ij} = \frac{A_i A_j}{\sigma^2} \sum_{n=-N}^{N} n^2 \cos[(\omega_i - \omega_j)n + (\phi_i - \phi_j)] \tag{8}$$

In particular, with $i = j$, the diagonal terms have the following simple close form expression,

$$J_{ii} = \frac{A_i^2}{\sigma^2} \cdot \frac{N(N+1)(2N+1)}{3} \tag{9}$$