Hindawi Publishing Corporation
EURASIP Journal on Information Security
Volume 2007, Article ID 75961, 12 pages
doi:10.1155/2007/75961

Research Article

Audio Watermarking through Deterministic plus Stochastic Signal Decomposition

Yi-Wen Liu[1, 2] **and Julius O. Smith**[1]

[1] *Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, Palo Alto, CA 94305, USA*
[2] *Boys Town National Research Hospital, 555 North 30th Street, Omaha, NE 68131, USA*

Correspondence should be addressed to Yi-Wen Liu, jacobliu@ccrma.stanford.edu

Received 1 May 2007; Revised 10 August 2007; Accepted 1 October 2007

Recommended by D. Kirovski

This paper describes an audio watermarking scheme based on sinusoidal signal modeling. To embed a watermark in an original signal (referred to as a *cover signal* hereafter), the following steps are taken. (a) A short-time Fourier transform is applied to the cover signal. (b) Prominent spectral peaks are identified and removed. (c) Their frequencies are subjected to quantization index modulation. (d) Quantized spectral peaks are added back to the spectrum. (e) Inverse Fourier transform and overlap-adding produce a watermarked signal. To decode the watermark, frequencies of prominent spectral peaks are estimated by quadratic interpolation on the magnitude spectrum. Afterwards, a maximum-likelihood procedure determines the binary value embedded in each frame. Results of testing against lossy compression, low- and highpass filtering, reverberation, and stereo-to-mono reduction are reported. A Hamming code is adopted to reduce the bit error rate (BER), and ways to improve sound quality are suggested as future research directions.

1. INTRODUCTION

The audio watermarking community has successfully adopted frequency-domain masking models standardized by MPEG. Below the masking threshold, a spread spectrum watermark (e.g., [1, 2]) distributes its energy, and the same threshold also sets a limit to the step size of quantization in informed watermarking [3]. Nevertheless, subthreshold perturbation is not the only way to generate perceptually similar sounds. Alternatively, a signal comprised of a large number of samples can be modeled with fewer variables called *parameters* [4]. Then, a watermark can be embedded in the signal through small perturbation in the parameters [5].

Audio signals can be parameterized while retaining surprisingly high sound quality. A classic parametric model is *linear prediction* [6], which enables speech to be encoded in filter coefficients and excitation source parameters [7]. Another model is to represent a tonal signal as a sparse sum of time-varying sinusoids [8, 9]. Although developed separately, predictive modeling and sinusoidal modeling have been used jointly [10]. A signal is modeled as a sum of sinusoids, and the residual signal that does not fit well to the model is parameterized by linear prediction. This hybrid system is referred to as being "deterministic plus stochastic" (D+S). The D component refers to the sinusoids, and the S component refers to the residual because it lacks tonal quality, therefore sounding like filtered noise. D+S decomposition was refined by Levine [11] by further decomposing the S component into a quasistationary "noise" part and a rapidly changing "transient" part. Levine's decomposition was given the name sines + noise + transients and considered as an efficient and expressive audio coding scheme. The development in D+S modeling has culminated in its endorsement by MPEG-4 as part of the audio coding standard [12].

In audio watermarking, meanwhile, the flexibility of D+S decompositions has brought forth a few novel schemes in recent years. Using Levine's terminology, watermarks have been embedded in two of the three signal components—in the transient part through onset time quantization, and in the sinusoids through phase quantization or frequency manipulation.

Embedding in the transients relies on an observation that the locations of a signal's clear onsets in its amplitude envelope are invariant to common signal processing operations

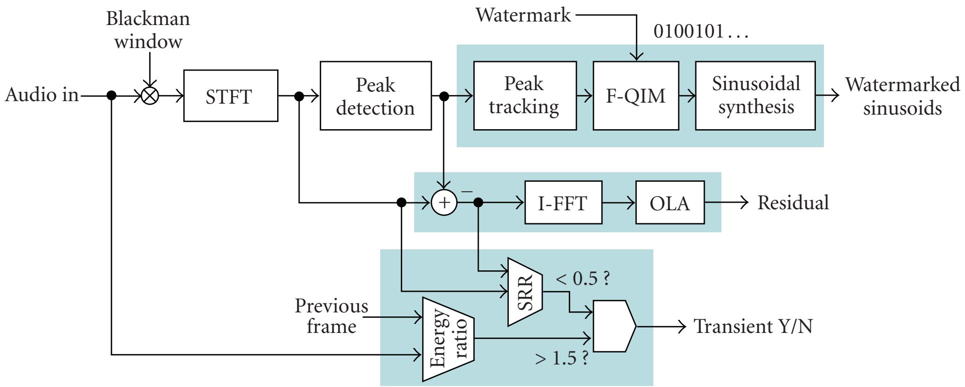


Figure 1: Signal decomposition and watermark embedding. Highlighted areas indicate (from top to bottom) the sinusoid processing modules, the residual computation modules, and the transient detection logic, respectively.

[13]. Such onsets, sometimes referred to as *salient points*, can be identified by wavelet decomposition [14] and quantized in time to embed watermarks; Mansour and Tewfik [15] reported robustness to MPEG compression (at 112 kbps/ch) and lowpass filtering (at 4 kHz), and their system sustained up to 4% of time-scaling modification with a probability of error less than 7%. Repetition codes were applied to achieve reliable data hiding at 5 bps (bits per second).

Phase quantization watermarking was first proposed by Bender et al. [16]. For each long segment of a cover signal, the phase at 32–128 frequency bins of the first short frame was replaced by $\pm\pi/2$, representing the binary 1 or 0, respectively. In all of the frames to follow, the relative phase relation was kept unchanged. More recently, Dong et al.[17] proposed a phase quantization scheme which assumes harmonic structure of speech signals. The absolute phase of each harmonic was modified by Chen and Wornell's quantization index modulation [18] (QIM) with a step size of $\pi/2$, $\pi/4$, or $\pi/8$. About 80 bps of data hiding was reported, robust to 80 kbps/ch of MP3 compression with a BER of approximately 1%.

Although phase quantization is shown as being robust to perceptual audio compression, human hearing is not highly sensitive to phase distortion, as argued by Bender et al. [16]. Thus, an attacker has the freedom to use imperceptible frequency modulation and steer the absolute phase of a component arbitrarily, thus defeating phase quantization schemes. Therefore, in the present work, we seek to embed a watermark not in the absolute phase of a component, but in its rate of change, the *instantaneous frequency.*

At first, audio watermarking by manipulating the cover signal's frequency was inspired by echo-hiding [16]. Petrovic [19] observed that an echo is a "replica" of the cover signal placed at a delay and the echo becomes transparent if it is sufficiently attenuated. He then attempted to place an attenuated replica at a shifted frequency to encode hidden information, but he did not disclose details of watermark decoding. Succeeding Petrovic's work, Shin et al. [20] utilized pitch scaling of up to 5% at mid frequency (3-4 kHz) for watermark embedding. Data hiding of 25 bps robust to 64 kbps/ch of audio compression were reported with BER <5%. A year later, we achieved 50 bps of data hiding by QIM in the frequency of sinusoidal models, but the algorithm only applied to synthetic sounds [5]. Independently, Girin and Marchand [21] studied frequency modulation for audio watermarking. In speech signals, surprisingly, frequency modulation in the 6th harmonic or above was found imperceptible up to a deviation of 0.5 times of the fundamental frequency. Based on this observation, transparent watermarking at 150 bps was achieved by coding 0 and 1 with positive and negative frequency deviations, respectively.

The watermarking scheme presented in this paper also induces frequency shifts to the cover signal but it differs from previous work in a few ways. First, the cover signal is *replaced by*, instead of being *superposed with*, the replica. This is achieved through sinusoidal modeling, spectral subtraction, and QIM in frequency (hereafter referred to as F-QIM). Second, the scale of frequency quantization, based on studies of pitch sensitivity in human hearing, is about an order of magnitude smaller than that described by Shin et al. [20] and Girin and Marchand [21]. The watermark decoding therefore requires unprecedented accuracy of frequency estimation. To this end, a frequency estimator that approaches the Cramér-Rao bound (CRB) is adopted. Third, as an extension to our previous work [5, 22], the new scheme is not limited to synthetic signals. Design of the new scheme is described next. Afterwards, in Section 3, robustness is evaluated, and results from a pilot listening test are reported. Rooms for improvement are pointed out in Section 4. Particularly, watermark security of the F-QIM scheme remains to be addressed. In this regard, this paper should be viewed as a proof of concept rather than a complete working solution.

2. METHODS

The watermark encoding process is based on the decomposition of a cover signal into sines + noise + transients. As shown in Figure 1, initially, the spectrum of the cover signal is computed by the short-time Fourier transform (STFT). If the current frame contains a sudden rise of energy and the sine-to-residual energy ratio (SRR) is low, it is labeled *transient* and passed to the output unaltered. Otherwise, prominent peaks are detected and represented by sinusoidal parameters. The *residual* component is computed by removing all the prominent peaks from the spectrum, transforming the spectrum back to the time domain through inverse FFT (I-FFT), and then overlap-adding (OLA) the frames in time. Parallel to this, a peak tracking unit memorizes sinusoidal parameters from the past and links peaks across frames to form trajectories. The watermark is embedded in the trajectories via QIM in frequency. The signal that takes quantized trajectories to synthesize consists of *watermarked sinusoids*. In this paper, a *watermarked signal* is defined as the sum of the watermarked sinusoids, the residual, and the unaltered transients. Details of each building block are described next.

2.1. Implementing D+S decomposition

Window selection

To compute STFT, the Blackman window [23] of length $L = 2N$ is adopted, $N = 1024$. Compared to the more commonly used Hann window, the Blackman window is better in terms of its side lobe rejection (57 versus 31 dB) and spectral roll-off rate (18 versus 12 dB per octave). Thus, the residual components after *spectral subtraction* (to be described) are masked better using the Blackman window.

Calculating the masking curve

Only unmasked peaks are used for watermark embedding. The masking curve is computed via a spreading function $\psi(z)$ that approximates the pure-tone excitation pattern on the human basilar membrane [24]:

$$\frac{d\psi}{dz} = \begin{cases} 0, & z_0 - 0.5 \le z \le z_0 + 0.5, \\ 27, & z < z_0 - 0.5, \\ -27, & z > z_0 + 0.5, \Lambda \le 40, \\ -27 + K(\Lambda - 40), & z > z_0 + 0.5, \Lambda > 40, \end{cases} \tag{1}$$

where Λ is the sound pressure level (SPL) in dB (re: 2×10^{-5} Pa), $K = 0.37, z_0$ is the pure tone's frequency in Barks [25], and z is the *critical band rate*, also in Barks, at other frequencies. $\psi(z_0) = 0$. Note that SPL is a physically measurable quantity. To align it with digital signals, a pure tone at the maximum amplitude (e.g., 1 for compatibility with MATLAB's wavread function) is arbitrarily set equal to 100 dB SPL. The masking level $M(z)$ is given by

$$M(z) = \Lambda - \Delta(z_0) + \psi(z), \tag{2}$$

where the offset $\Delta = (14.5 + z_0)$ dB [26].[1]

To express $M(z)$ in units of power per frequency bin, the following normalization is necessary [28]:

$$M_k^2 = \frac{10^{M(z)/10}}{N(z)}, \tag{3}$$

where $N(z)$ is the equivalent number of FFT bins within a critical bandwidth (CBW) [25] centered at $z = z(k\Omega)$, with $k\Omega = \mathrm{k}(2\pi/N_{\mathrm{FFT}})$ being the frequency of the kth bin.

When more than one tone is present, the overall masking curve $\sigma^2(k\Omega)$ is set as the maximum of the spreading functions and the threshold in quiet $I_0(f)$:

$$\sigma^2(k\Omega) = \max\{M_{1,k}^2, M_{2,k}^2, \ldots, M_{j,k}^2, 10^{I_0(k\Omega)/10}\}, \tag{4}$$

where $M_{j,k}$ denote the masking level at frequency bin k due to the presence of tone j, and $I_0(f)$ is calculated using Terhardt's approximation [29]:

$$I_0(f)/\mathrm{dB} = 3.64 f^{-0.8} - 6.5 e^{-0.6(f-3.3)^2} + 10^{-3} f^4, \tag{5}$$

where$f$ is in the unit of kHz. In this paper, a peak is considered "prominent" if its intensity is higher than the masking curve. To carry a watermark, prominent peaks will be subtracted from the spectrum and then added back at quantized frequencies.

Spectral interpolation and subtraction

Sinusoidal modeling parameters are estimated via a quadratic interpolation of the log-magnitude FFT (QIFFT) [30]. Blackman windowed signals of length 2048 are first zero-padded to a length of 2^{14} before FFT. Denote the 2^{14}-length discrete spectrum $S_k = S(k\Omega), \Omega = 2\pi/2^{14}$. Any peak such that$|S_k| > |S_{k+1}|$ and $|S_k| > |S_{k-1}|$ is associated with frequency and amplitude estimates given by

$$\begin{aligned} \widehat{\omega} &= \left(k + \frac{1}{2}\frac{a^- - a^+}{a^- - 2a + a^+}\right)\Omega, \\ \log \widehat{A} &= a - \frac{1}{4}\left(\frac{\widehat{\omega}}{\Omega} - k\right)(a^- - a^+) - C, \end{aligned} \tag{6}$$

where $a^- = \log|S_{k-1}|$, $a^+ = \log|S_{k+1}|$, $a = \log|S_k|$, and $C = \log(\sum_{n=-N}^{N} w^{\mathrm{B}}[n])$ are a normalization factor, with $w^{\mathrm{B}}[n]$ being the Blackman window. Denote $q = (\widehat{\omega}/\Omega) - k$. The phase estimate is given by linear interpolation:

$$\widehat{\phi} = \angle S_k + q(\angle S_{k+1} - \angle S_k). \tag{7}$$

The sinusoid parameterized with $\{\widehat{A}, \widehat{\omega}, \widehat{\phi}\}$ can be removed by *spectral subtraction*, as described below.

Step 0. Initialize the sum spectrum $\widehat{S}(\omega) = 0$ and denote $\widehat{S}_k = \widehat{S}(k\Omega)$.

[1] The spreading function in (1) is similar to MPEG psychoacoustic model 1 (in ISO/IEC 11172-3). They share a few common features. First, the spreading function rolls off faster on the low-frequency side than on the high-frequency side. Second, the slope on the high-frequency side decreases as the sound level increases. However, what this psychoacoustic model lacks is the ability to differentiate between tonal and nontonal maskers so as to set $\Delta(z_0)$ accordingly. In (2), this model always assumes that maskers are tonal. Readers interested in calculation of a tonal index can refer to [27, Chapter 11].

Step 1. For each peak, fit the main lobe of the Blackman window transform $W(\omega)$ at $\widehat{\omega}$, scale it by $\widehat{A}\exp(j\widehat{\phi})$,[2] and denote the scaled and shifted main lobe of the window as

$$\widehat{W}(\omega) = \begin{cases} \widehat{A}e^{j\widehat{\phi}}W(\omega - \widehat{\omega}) & \text{if } |\omega - \widehat{\omega}| \le 3\dfrac{2\pi}{L}, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Step 2. Denote $\widehat{W}_k = \widehat{W}(k\Omega)$ and update $\widehat{S}_k$ by $\widehat{S}_k + \widehat{W}_k$.

Step 3. Take the next prominent peak and repeat steps 1 and 2 until all prominent peaks are processed; and towards the end, $\widehat{S}_k$ becomes the spectrum to be subtracted.

Step 4. Define the residual spectrum R_k as follows:

$$R_k = \begin{cases} S_k - \widehat{S}_k & \text{if } |S_k - \widehat{S}_k| < |S_k|, \\ S_k, & \text{otherwise.} \end{cases} \tag{9}$$

The if condition in (9) guarantees that the residual spectrum is smaller than the signal spectrum everywhere, in terms of its magnitude.

2.2. Residual and transient processing

Inaudible portion of the residual is removed by setting R_k to zero if $|R_k|^2$ is below the masking curve. Then, inverse FFT is applied to obtain a residual signal $\mathbf{r}$ of the length N_{FFT}. Due to concerns that will be discussed later regarding perfect reconstruction, $\mathbf{r}$ is shaped in the time domain according to

$$r^{\text{sh}}[n] = r[n]\left(\frac{w^{\text{H}}[n]}{w^{\text{B}}[n]}\right), \tag{10}$$

where $w^{\text{H}}[n]$ denotes Hann window of length N. Then, across frames, $r^{\text{sh}}[n]$ is overlap-added with a hop of length $h = N/2$ to form the final residual signal $r^{\text{OLA}}[n]$:

$$r^{\text{OLA}}[n] = \sum_{m=1}^{\infty} r_m^{\text{sh}}[n - mh], \tag{11}$$

where the subscript m is an index pointing to the frame centered around time $n = mh$.

Regions of rapid transients need to be identified and treated with caution so as to avoid *pre-echoes,* which occur when the short-time phase spectrum of a rapid onset is modified. If a pre-echo extends beyond the range of the onset's backward masking [25], it becomes an audible artifact. To avoid pre-echoes, in the current study, regions of rapid onsets are kept unaltered. A frame is labeled "transient" if all of the following conditions are true.

(i) The sines-to-residual energy ratio in the current frame is less than 5.0.

(ii) The energy ratio of the current frame to the previous frame is greater than 1.5.

(iii) There is at least a peak greater than 30 dB SPL between 2 and 8 kHz.

When all three criteria are met, spectral subtraction and watermark embedding are disabled for 2048 samples around the current frame. The signal fades in and out of the transient region using Hann window of length 1024 with 50% overlap.

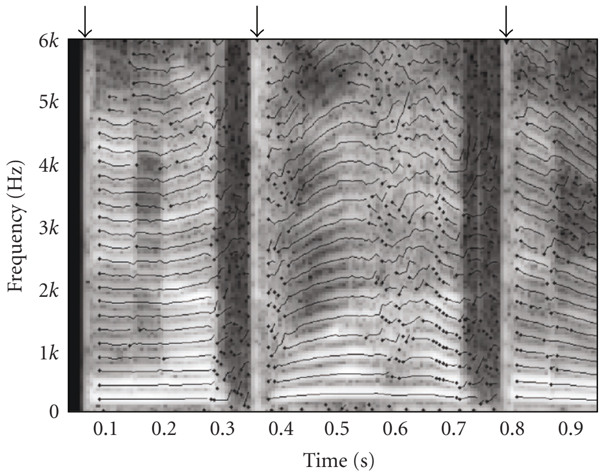


Figure 2: Frequency trajectories extracted from a recording of German female speech, overlaid on its spectrogram. Onsets of trajectories are marked with dots. Arrows point to transient regions, where peak detection is temporarily disabled.

2.3. Watermarking the sinusoids

Peak tracking

Denote the estimated frequencies of the peaks as $\{\omega'_j\}$ and $\{\omega_j\}$ at previous and current frames, respectively. The following procedure connects peaks across the frame boundary.

Step 1. For each peak j in the current frame, find its closest neighbor $i(j)$ from the previous frame; $i(j) = \arg\min_k |\omega'_k - \omega_j|$, and connect peak $i(j)$ of the previous frame to peak j of the current frame.

Step 2. If a connection has a frequency slope greater than 20 barks per second, break the connection and label peak j of the current frame as an onset to a new trajectory.

Step 3. If a peak i_0 in the previous frame is connected to more than one peak in the current frame, keep only the connection with the smallest frequency jump, and mark all the other peaks j such that $i(j) = i_0$ as onsets to new trajectories.

A trajectory starts at an onset and ends whenever the connection cannot continue. Trajectories extracted from a recording of German female speech are shown in Figure 2.

[2] For convenience of discussion, assume that the normalization factor is $C = 0$.

Sinusoidal synthesis

For each trajectory k, let $\phi_0^{(k)}$ denote the initial phase, $\{A_{km}\}$ its amplitude envelope, and $\{\omega_{km}\}$ its frequency envelope. A window-based synthesis can be written as

$$s_{\text{total}}[n] = \sum_k \sum_m A_{km} w[n - mh] \cos\left(\phi_m^{(k)} + \omega_{km}(n - mh)\right), \tag{12}$$

where the phase $\phi_m^{(k)}$ is updated as follows:

$$\phi_m^{(k)} = \phi_{m-1}^{(k)} + \left(\frac{\omega_{k,m-1} + \omega_{km}}{2}\right) h. \tag{13}$$

In (12), the window $w[n]$ needs to satisfy a perfect reconstruction condition

$$\sum_{m=-\infty}^{\infty} w[n - mh] = 1 \quad \forall\, n. \tag{14}$$

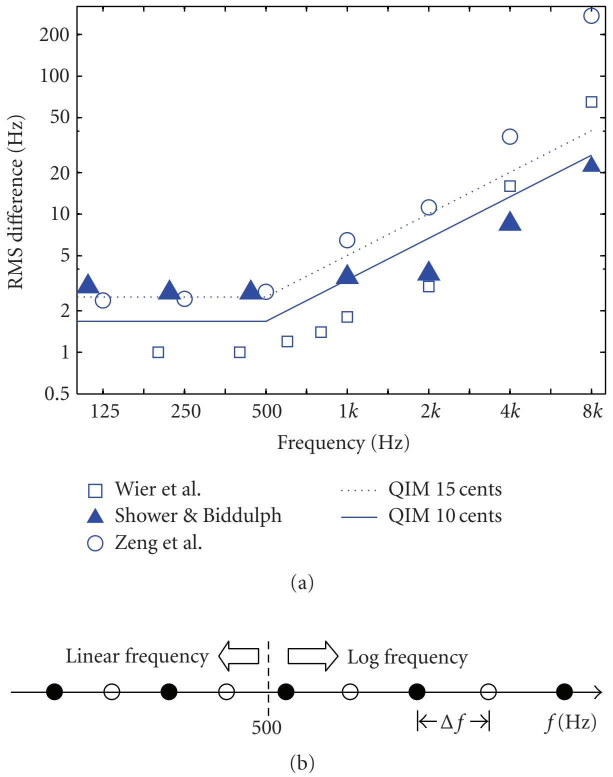
To be consistent with residual postprocessing in (10), the Hann window is adopted in (12).

Designing frequency quantization codebooks

Frequency parameters $\{\omega_{km}\}$ in (12) are quantized to embed a watermark. The just noticeable difference in frequency, or *frequency limen* (FL), is considered in the design of the quantization codebooks. Figure 3(a) shows existing measurements of the FL from human subjects with normal hearing [31–33]. Levine [11] reported that a sufficiently small frequency quantization at approximately a fixed fraction of a CBW did not introduce audible distortion. This design is adopted in the sense that the frequency quantization step size Δf is a constant below 500 Hz and linearly increases above 500 Hz (see Figure 3(b)). The root-mean-square (RMS) frequency shift incurred by F-QIM is plotted in Figure 3(a) for comparison.

Repetition coding schemes

In principle, one bit of information can be embedded in every prominent peak at every frame. Liu and Smith [22] demonstrated over 400 bps of data hiding in a synthesized signal that has 8 well-resolved sinusoidal trajectories throughout its whole duration. However, for recorded signals, sinusoids are not as stationary and well resolved. Therefore, in the current study, two repetition-coding schemes are adopted to reduce the BER at the cost of lowering the data-hiding payload. First, in each frame, all prominent peaks are frequency-aligned to either one set of QIM grid points or the other, thus reducing the data-hiding rate to one bit per frame. Second, adjacent frames are pairwise enforced to have identical peak frequencies so as to produce sinusoids that perfectly align to QIM grid points at every other hop of length h. This simplifies watermark decoding, but it might degrade sound fidelity. More careful study of the sound quality is left for future investigation. Hereafter, the data-hiding payload is set at one bit per $2h$ samples unless otherwise mentioned. At a 44.1 kHz sampling rate, this data-hiding payload is approximately 43 bps.



(a)

(b)

FIGURE 3: Quantization step size and just noticeable difference in frequency. (a) Behavioral measurement of FL. The stimuli used by Wier et al. [32] were pure tones; the stimuli in Shower and Biddulph [31] were frequency-modulated tones. (b) Design of the F-QIM codebooks. Open and filled circles represent the two binary indexes, respectively. The step size is approximately a fixed fraction of the CBW.

2.4. Watermark decoding

Frequency estimation

To decode a watermark, frequencies of prominent spectral peaks are estimated using the Hann window of length h. It is desired that the frequency estimation is not biased and that the error is minimized. Abe and Smith [30] showed that the QIFFT method efficiently achieves both goals to a perceptual accurate degree if, first, the spectrum is sufficiently interpolated, second, the peaks are sufficiently well separated, and third, the SNR is sufficiently high. When only one peak is present, zero-padding to a length of $5h$ confines frequency estimation bias to $10^{-4}F_s/h$. If multiple peaks are present but separated by at least $2.28F_s/h$, the frequency estimation bias is bounded below $0.042F_s/h$. If peaks are well separated and SNR is greater than 20 dB, then the mean-square frequency estimation error decreases as SNR increases. The error either approaches the CRB (at moderate SNR) or is negligible compared to the bias (at very high SNR). In all experiments to be reported in the next section, the QIFFT method was adopted as the frequency estimator at the decoder; the windowed signal is zero-padded to the length $8h$.

Maximum-likelihood combination of "opinions"

When the watermark decoder receives a signal and identifies peaks at frequencies $\{\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_J\}$, these frequencies are decoded to a binary vector $\mathbf{b} = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_J)$ with error probabilities $\{p_j\}$. To determine the binary value of the hidden bit while some $\hat{b}_j$'s are zeros and some are ones, the following hypothesis test is adopted:

$$b^{\text{opt}} = \begin{cases} 1 & \text{if} \sum_{j=1}^{J} \left[\log\left(\frac{1-P_j}{P_j}\right)\right]\left(\hat{b}_j - \frac{1}{2}\right) > 0, \\ 0 & \text{otherwise}. \end{cases} \qquad (15)$$

Equation (15) is a maximum-likelihood (ML) estimator if bit errors occur independently and the prior distribution is $p(0) = p(1) = 0.5$. Note that the error probabilities $\{P_j\}$ are not known a priori. If we assume that the frequency estimation error (FEE) is normally distributed, not biased, and its standard deviation is equal to the CRB, then let us approximate P_j by the probability that the absolute FEE exceeds half of QIM step size:

$$P_j \approx 2Q\left(\frac{\Delta f_j/2}{J_{ff}^{-1/2}}\right), \qquad (16)$$

where $Q(x) = (1/\sqrt{2\pi})\int_x^\infty e^{-u^2/2}du$, Δf_j is the QIM step size near f_j, and $J_{ff}^{-1/2}$ denotes the CRB for frequency estimation. Note that the CRB depends on how the attack on the watermark is modeled. Currently, the system simply assumes that the attack is additive Gaussian noise. Therefore [34, 35],

$$J_{ff} = \left(\frac{\partial \mathbf{S}}{\partial f_j}\right)^{\dagger} \overset{-1}{\mathbf{\Sigma}} \left(\frac{\partial \mathbf{S}}{\partial f_j}\right), \qquad (17)$$

where $\mathbf{S}$ represents the DFT of the signal $S_{\text{total}}[n]$ defined in (12), and $\mathbf{\Sigma}$ is the power spectral density of the additive Gaussian noise. In all the experiments to be reported next, the noise spectrum $\mathbf{\Sigma}$, unknown to the decoder a priori, is taken as the maximum of the masking curve in (4) and the residual magnitude in (9).[3]

3. EXPERIMENTS

In this section, a previous report on the performance of F-QIM watermarks is summarized. Then, results obtained from a new set of music samples are presented, including robustness and sound-quality evaluation.

3.1. Watermarking sound quality assessment materials

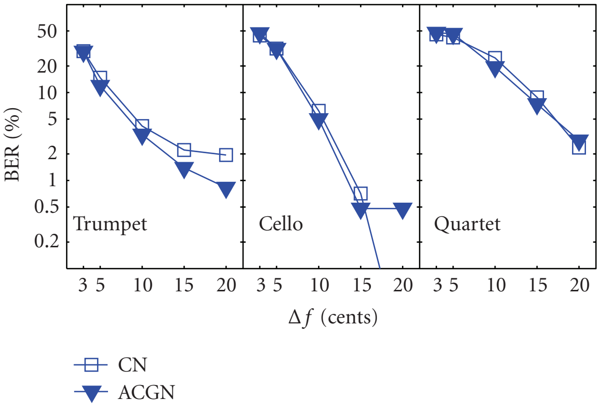


FIGURE 4: Noise robustness of F-QIM watermarking.

In our previous study [34], two types of noise were introduced to single-channel watermarked signals as a preliminary test of robustness. The cover signals are selected from the European Broadcast Union's *sound quality assessment materials* (EBU SQAM).[4] BER was measured as a function of the F-QIM step sizes between 3 and 20 cents (at $f > 500$ Hz). The first type of noise is additive colored Gaussian noise (ACGN). The ACGN's SPL was set at the masking threshold at every frequency. The second type of noise was the coding noise (CN) imposed by variable-rate compression using the open-source perceptual audio coder *Ogg Vorbis* (available at www.vorbis.com).

Results from three soundtracks are shown in Figure 4. Unsurprisingly, the watermark decoding accuracy increases as a function of the quantization step size. Given the performance shown in Figure 4, it becomes crucial to find the F-QIM step size that has an acceptable BER and does not introduce objectionable artifacts. Informal listening tests by the authors suggested that human tolerance to F-QIM depends on the timbre of the cover signal. For example, sinusoids in the trumpet soundtrack are quite stationary whereas other soundtracks may have higher magnitudes of vibrato. Therefore, a smaller F-QIM step size was necessary for the trumpet soundtrack. This finding is consistent with the fact that the FL is larger for FM tones than for pure tones, as shown in Figure 3.

To this date, choosing the F-QIM step size adaptively remains a future goal. The step size was picked at {5, 10, 15} cents for {trumpet, cello, quartet} soundtracks, respectively. Thus, BER was {12%, 5%, 7%} against ACGN and {15%, 6%, 9%} against CN. Also, on average, BER was about 13% against lowpass filtering at a cutoff frequency of 6 kHz, 19% against 10 Hz of full-range amplitude modulation, and 24% against playback speed variation. However, the F-QIM watermarks failed to sustain pitch scaling beyond half of the quantization step size and were vulnerable to desynchronization in time. A detailed report can be found in [34].

[3] The cover signal remains unknown to the decoder; the masking curve and the residual are computed entirely based on the received signal.

[4] They are available at http://sound.media.mit.edu/mpeg4/audio/sqam/, as of March 5, 2007.

Table 1: Music selected in experiment 3.2. The last two columns show BERs when decoding directly from the watermarked signal.

No.	Label	Sound description	Genre	BER (%)	
				Ch1	Ch2
1	*Smetana*	Excerpt from the symphonic poem *Má Vlast: the Moldau*	Instrumental	10.7	13.6
2	*Brahms*	Piano quartet op. 25; opening part of the 4th movement: *Presto*	Instrumental	13.7	15.1
3	*Frère Jacques*	French song, with bells in the background	Vocal	18.1	15.3
4	*Il Court le Furet*	French song, with sounds of percussion and electronic keyboard in the background	Vocal	6.5	7.7
5	*Christian Pop I*	*Thank You for Giving to the Lord;* Contemporary American Christian song, featuring a tenor voice	Vocal	10	11.4
6	*Chrisitan Pop II*	Another excerpt from of the same song	Vocal	12.5	16.8
7	*Señora Santana*	Spanish song, featuring a duet sung by two girls and accompanied by piano, guitar, and percussions	Vocal	6.5	7.1
8	*El Coquí*	Spanish song of Puerto Rican origin, accompanied with pipe-flute, guitar, bass, and percussion	Vocal	14.0	12.5
9	*Ella Fitzgerald I*	*I'm Gonna Go Fishing*; alto voice accompanied by a jazz band	Vocal	5.4	4.9
10	*Ella Fitzgerald II*	*I Only Have Eyes for You;* jazz band introduction and alto voice entrance	Vocal	9.6	11.4
11	*Liszt I*	Piano entrance, a slow arpeggio, accompanied by the string section (the following four samples are from Liszt's Piano Concerto no. 2)	Instrumental	32.4	28.3
12	*Liszt II*	Piano and horn duet	Instrumental	27.7	22.5
13	*Liszt III*	Mostly piano solo, featuring a long descending semitonal scale	Instrumental	14.1	11.8
14	*Liszt IV*	Finale: piano plus all sorts of instruments in the orchestra	Instrumental	18.9	14.8
15	*Stravinsky I*	Opening part of the 1st movement in *Trois Mouvements de Petrouchka,* featuring fast piano solo with much staccato	Instrumental	11.9	11.8
16	*Stravinsky II*	From the 2nd of the *Three Movements*, featuring slow piano solo with phrases in legato	Instrumental	10.5	9.3
17	*Bumble Bee*	Rimsky-Korsakov's *Flight of the Bumble Bee,* featuring cellist Yo-Yo Ma and singer Bobby McFerrin	Voice as an instrument	17.1	15.0
18	*Ave Maria*	McFerrin on Bach's prelude line and Ma on Gounod's Ave Maria rendition	Voice as an instrument	6.3	7.2
			Average	13.7 ± 7.2	13.1 ± 5.6

3.2. Watermarking stereo music

To test the system further, watermarks are embedded in 18 sound files, each 20 seconds long. All the files are stereo recordings in standard CD format (44.1 kHz sampling rate, 16-bit PCM) from Yi-Wen Liu's own collection of CDs. Brief description of the music can be found in Table 1.

The F-QIM step size is 12 cents above 500 Hz, the same for all files. The attempted data-hiding rate is 43 bps. The watermarking scheme is evaluated in terms of its robustness to the following procedures.

(1) *Lowpass filtering (LPF).* Lowpass finite impulse response (FIR) filters of length 65 are obtained by Hamming windowing of the ideal lowpass responses. The cutoff frequency is 4–10 kHz.

(2) *Highpass filtering (HPF).* Highpass FIR filters of length 65 are obtained using MATLAB's fir1 function. The cutoff frequency is 1–6 kHz.

(3) *MPEG advanced audio coding (AAC).* Stereo watermarked signals are compressed and then decoded using *Nero Digital Audio*'s high-efficiency AAC codec (HE-AAC) [36]. The compression bit rate is constant at 80, 96, 112, or 128 kbps/stereo (i.e., 40–64 kbps/ch).

(4) *Reverberation (RVB).* Room reverberation is simulated using the image method [37]. The dimensions of the virtual room and the locations of the sources and microphone are shown in Figure 5. For convenience of discussion, the reflectance R is set equally on the walls, ceiling, and floor. To compute the impulse response from one source to the microphone, 24 reflections are considered along each of the 3 dimensions, resulting in 25^3 coupling paths. The impulse response is then convolved with the watermarked signal.

(5) *Reverberation plus stereo-to-mono reduction (RVB + S/M).* To simulate mono reduction, both sound sources in the virtual room are considered. An identical bit stream is embedded in both channels of the stereo signal. The two channels of the watermarked signal are simultaneously played at the two virtual source locations, respectively. A mono signal is virtually recorded at the microphone location using the image method with reflectance $R = 0.6$.

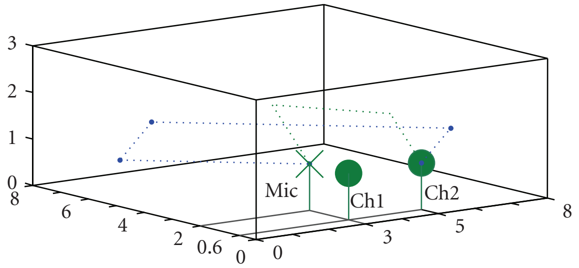
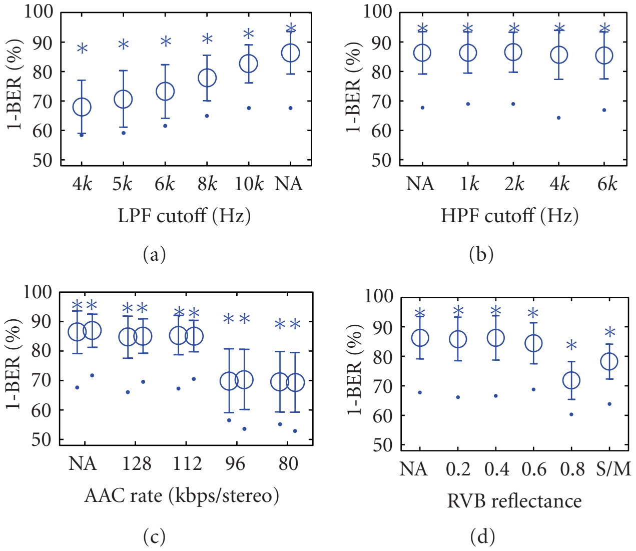


Figure 5: Configuration of the virtual recording room (8m × 3m × 3m). Circles indicate the locations of the two loudspeakers. The microphone and the two loudspeakers are at the same height (1m). Two possible coupling paths from channel 2 to the microphone are illustrated, each bouncing off the walls a few times. Sounds are also allowed to reflect from ceiling and floor.



(a) (b) (c) (d)

Figure 6: Performance of F-QIM watermarking scheme against LPF, HPF, AAC, and RVB (+ S/M). NA = no attack. Circles and error bars indicate mean ± standard deviation across 18 files. Dots and asterisks indicate the worst and the best performances among 18 files, respectively. For AAC, results from both channels are shown separately. For other types of attacks (except RVB + S/M), results from ch1 are shown.

Figure 6 shows BER at different levels of signal processing. The top left panel shows a gradual loss of performance against LPF as the cutoff frequency decreases. However, as shown on the top right panel, the performance seems to sustain HPF even when the watermarked signals are cut off below 6 kHz.

At 112 kbps/stereo, performance against AAC is comparable to direct decoding without attack. However, it drops abruptly when the signals are compressed to 96 kbps/stereo. Similarly, performance remains good at mid to low levels of reverberation ($R \leq 0.6$), but it drops significantly at $R = 0.8$.

As shown on the lower right panel, at $R = 0.6$, adding ch2 causes about 6% more errors than virtual recording solely with ch1.

3.3. PEAQ-anchored subjective listening test

To evaluate the sound quality of watermarked signals, 14 subjects were recruited for a pilot listening test. The goal of this test was to tell whether watermarked signals sound better or worse than their originals plus white noise.[5] The test consists of three modules. Each module contains an audio file R = the reference (in wav format) from Table 1, and three other files. One of the three files is identical to R, one is watermarked (WM), and one is R plus Gaussian white noise (R+WN). The subjects did not know beforehand the identity of the three files, and the three files were given random names that did not reveal their identities. Subjects were asked to find a good listening device and a quiet place so as to identify the file that is identical to R by ears. There was no time limit; subjects could repeatedly listen to all the files. Additionally, they were asked two questions regarding the remaining two files.

(1) Which one's distortion is more noticeable?

(2) Which one is more annoying?

The noise levels in R+WN signals were carefully chosen so that their objective difference score (ODG), as computed by PEAQ (*Perceptual Evaluation of Audio Quality*, ITU-R BS.1387) [38], had a reasonable range for a comparative study (Table 2, last two columns). Note that ODG = −1 infers that the difference to the reference file is noticeable but not annoying, −2 infers that the difference is somewhat annoying, −3 annoying, and −4 very annoying.

This group of subjects did not always identify R accurately (Table 2, second column). One subject had wrong answers in all three test modules, so his response is excluded in the following analyses. Of all the other wrong answers, WMs were misidentified as R for six times; only once was R+WN mistaken as R. Regarding clips nos. 2 and 8, a definite majority of subjects who correctly identified R said that WM sounded better than R+WN (Table 2, 3rd column). Mixed results were obtained for clip no. 18.[6] Assuming that the ODGs of R+WN were reliable, these results suggest that these subjects, as a group, would have rated the WM signals as better than annoying (clip no. 2), better than somewhat annoying (no. 8), or nearly somewhat annoying (no. 18).

Among the 14 subjects, 10 are active musicians (playing at least one instrument or voice), including three audio/speech engineers, three music researchers in the academia, and two composers.

[5] We knew that the F-QIM scheme does not achieve complete transparency yet. It would be nice if the sound quality can be evaluated objectively. However, known standards such as ITU-R BS.1387 are highly tuned to judge the artifacts introduced by compression codecs. They are not suitable to judge sinusoidal models. Therefore, we designed this alternative way to evaluate the quality of watermarked signals by comparing them to noise-added signals, which can be graded fairly by objective measures.

[6] All but one subject reported that the more noticeable distortion was always more annoying. One particular subject commented that white noise was more noticeable but easy to ignore. She reported that she could tolerate the WM in clip no. 2, but not in no. 18. She also said that WM in clip no. 8 was hard to distinguish from the reference. Based on her anecdotes, her preference was counted in favor of WM for clips nos. 2 and 8, and in favor of R+WN for clip no.18.

Table 2: PEAQ-anchored listening test. C: number of correct answers. M: number of times WM was misidentified as R. N: number of times R+WN was misidentified as R. Φ: number of subjects who admitted that they could not tell.

Reference signal	Accuracy in identifying R	Subjects' preference		Noise level (dB SPL)	ODG of R+WN
	(C : M : N : Φ)	WM	R+WN		
No. 2 (*Brahms*)	10 : 2 : 0 : 1	9	1	44	−2.6
No. 8 (*El Coquí*)	8 : 2 : 0 : 3	8	0	54	−2.1
No. 18 (*Ave Maria*)	8 : 2 : 1 : 2	4	4	34	−1.8

4. DISCUSSION

4.1. Robustness

Among the results reported in Figure 6, note that the watermarks withstood HPF but not LPF. This indicates that the system, as it is currently implemented, relies heavily on high-frequency (>6 kHz) prominent peaks. Therefore, when a signal processing procedure fails to preserve high-frequency peaks, the watermark's BER can significantly increase. For example, the mean BER nearly doubles (from 13.7% to 27.6%) at 6 kHz LPF.

Dependence on high-frequency sinusoids can also explain the sudden increase of BER when the AAC compression rate drops below 112 kbps/stereo. When available bits in the pool are not sufficient to code the sound transparently, the HE-AAC encoder either introduces LPF or switches to *spectral band replication* (SBR) [36] at high frequencies to ensure overall optimal sound quality. In the latter case, components at high frequency are parameterized by spectral envelopes. Peak frequencies can be significantly changed so that they foil the current implementation of F-QIM watermarking. This being said, however, the exact causes of degraded watermark performance at 96 kbps/stereo are worth of further investigation.

As shown in Table 1 and Figure 6, the watermark embedded by 12 cents of F-QIM shows widely different levels of robustness in different sound files. In general, with BER = 10–30%, error correction coding is necessary before F-QIM and can be adopted in various applications. A pilot study on repetition coding and error correction has been conducted, and the results are shown next.

4.2. Repetition coding and error correction

Clips nos. 11, 12, 14, and 17, whose BERs were among the worst (15–33%, Table 1), were chosen as the test bench. To hide a binary message, the message was first encoded with a Hamming(7,4) code (see, e.g., [39]). The Hamming code consists of $2^4 = 16$ code words of length 7, and up to 1 bit of error in every word can be corrected. Then, the resulting binary sequence went through repetition coding, and the output modulated the frequency quantization index at the frame rate $\doteq$ 43 bps.

Two different repetition coding strategies called, respectively, bit- and block-repeating were tested. The first strategy repeats each bit consecutively. For instance, {001 ...} becomes {000 000 111 ...} if the repetition factor $r = 3$. The second strategy repeats the whole input sequence. For instance, {1000011 ...} becomes {1000011 ... 1000011 ... 1000011 ...} if $r = 3$. For the second strategy to work, the encoder has to know the length of music in advance, and the hidden message should not be retrieved until the last repetition block is decoded. Nevertheless, the block-repeating strategy has an advantage. It is more effective in reducing the BER if decoding errors tend to occur in adjacent bits. This is clearly what we found empirically. In Figure 7, block repetition strategy (left panel, diamonds) consistently performed better than bit repetition (dots). Results from different files are color-coded, with blue = clip 11, ch1; green = clip 12, ch2; orange = clip 14, ch1; red = clip 17, ch1.

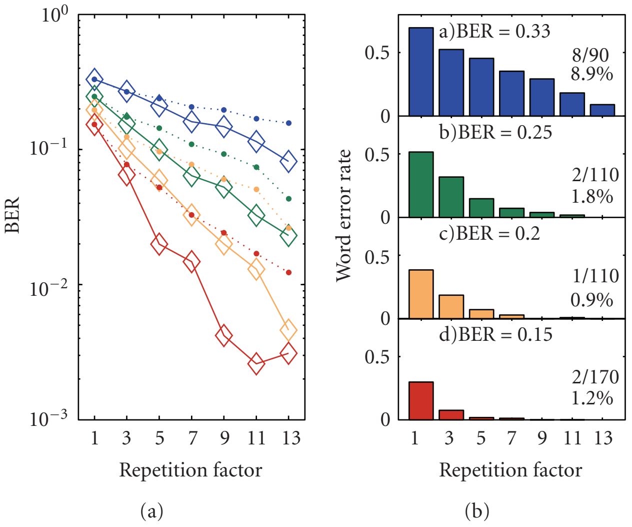


Figure 7: Effectiveness of repetition coding and error correction. (a) Decoding BER before error correction. (b) Wordwise decoding error rate using the block-repeating strategy and Hamming error correction. BERs listed here are as obtained before repetition coding and error correction.

In Figure 7, every data point is an average of 10 attempts using randomized hidden message. Empirically, when the raw BER ≤0.25, the block repetition strategy was able to reduce the error rate to <4% at $r = 13$, which led to zero error after Hamming correction. At a raw BER = 0.33, however, this coding scheme produced 8 word errors out of 90 trials.

With $r = 13$, the data payload is (20 sec) × (43 bps)/13 × 4/7 = 36 bits. In the future, if BER can be confined to <25% under common signal processing procedures, F-QIM should be useful for nonsecure applications. For applications with more stringent security requirements, a private key would need to be shared by the encoder and the decoder so the repetition code is pseudorandomized.

4.3. Other suggestions for future research

To improve the performances against LPF, one can adopt a multirate sinusoidal model [11] for watermark embedding. At low frequency, a longer window can be used in D+S signal decomposition to produce higher accuracy in frequency estimation. In this case, the data-hiding payload is reduced to trade for enhanced robustness. At high frequency, the watermark encoding configuration can remain the same inasmuch as to sustain HPF and high-quality AAC encoding.[7]

The virtual room experiments (see Figure 5) can be regarded as a pilot study of robustness against the playback-recording attack. The system currently shows an increase in BER when the reflectivity of the virtual room increases above $R = 0.6$. Thus, the system is robust to echoes up to $R = 0.6$ in this room. It is promising that the increase in BER is manageable in stereo-to-mono recording. However, note that the distances between {ch1, ch2} and the microphone are carefully chosen to avoid desynchronization. The delays are about 4.1 and 7.1 milliseconds from the two channels, or 180 samples and 312 samples (at $F_s = 44.1\,\text{kHz}$), which are shorter than the window length $h = 512$ at the decoder.

To provide a mechanism of self-synchronization, in the future, derived features from the trajectories could be chosen as the watermark-embedding parameters. Higher-dimensional quantization lattices, such as the spread transform scalar Costa scheme [40] and vector QIM codes [41], are worth of investigation. At the system level, an alternative approach is to embed another watermark in the transient part to provide synchronization in time (e.g., [13, 15]). The watermark carried by the deterministic components can thus be recovered using synchronization information from the transients' watermark. This could be interesting for broadcast monitoring applications, and we foresee little conflict in simultaneously embedding the two watermarks because the sinusoidal and transient components are decoupled in time.

In addition to watermarks embedded in tonal frequency trajectories and transients, the "noise" component of a sines + noise + transients model might be utilized for watermarking as well. To our knowledge, this has not been reported previously although spread spectrum watermarking methods are obviously closely related. A "noise" watermark and F-QIM watermark may mutually interfere since they overlap in both time and frequency. A noise-component watermark cannot be expected to survive perceptual audio coding schemes as well as tonal and transient watermarks. However, watermarks based on high-level features of the noise component, such as overall bandwidth variations, power envelope versus time, and other spectral feature variations over time, should survive audio coding well enough, provided that preservation of the chosen features is required for good audio fidelity.

Table 3: List of constants and frequently used symbols.

Symbol	Meaning	Default Value
F_s	Sampling rate	44.1 kHz
L	Blackman window length	2048
N	Hann window length	$L/2$
h	Hop size for sinusoidal synthesis	$N/2$
N_{FFT}	FFT length after zero-padding	8 L at encoder; 8 h at decoder
i, j, k, m	Dummy indexes, with an exception that j can also refer to the square root of -1 when there is no confusion	—
n	Discrete time index	—
A	Linear amplitude	—
f	Frequency in Hz	—
ω	Frequency in rad/sample	—
ϕ	Phase	—
Δf	Frequency quantization step size	—

Finally, the listening test results suggest that there is still room to diagnose the cause of artifacts, to modify the signal decomposing methods, and hence to improve the sound qualities. It is very important for an audio watermarking scheme to maximally preserve sound fidelity. To conclude, audio watermarking through D+S signal decomposition is still in its infancy, and many open ideas remain to be explored.

ACKNOWLEDGMENTS

The authors would like to thank the editors for encouraging words and two anonymous reviewers for highly constructive critiques. They also thank all friends who volunteered to take the listening test and provided valuable feedback.

REFERENCES

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

[2] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," *Signal Processing*, vol. 66, no. 3, pp. 337–355, 1998.

[3] J. Chou, K. Ramchandran, and A. Ortega, "Next generation techniques for robust and imperceptible audio data hiding," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '01)*, vol. 3, pp. 1349–1352, Salt Lake City, Utah, USA, May 2001.

[4] B. L. Vercoe, W. G. Gardner, and E. D. Scheirer, "Structured audio: creation, transmission, and rendering of parametric sound representations," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 922–939, 1998.

[5] Y.-W. Liu and J. O. Smith, "Watermarking parametric representations for synthetic audio," in *Proceedings IEEE Interna-*

[7] According to Apple Inc., "AAC compressed audio at 128 Kbps (stereo) has been judged by expert listeners to be 'indistinguishable' from the original uncompressed audio source." (See http://www.apple.com/quicktime/technologies/aac/ for more information.)

tional Conference on Acoustics, Speech and Signal Processing (ICASSP '03), vol. 5, pp. 660–663, Hong Kong, April 2003.

[6] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, Springer, New York, NY, USA, 1976.

[7] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '85)*, vol. 10, pp. 937–940, Tampa, Fla, USA, April 1985.

[8] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transaction Acoustics, Speech, Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[9] J. O. Smith and X. Serra, "PARSHL: an analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in *Proceedings of the International Computer Music Conference (ICMC '87)*, pp. 290–297, Tokyo, Japan, 1987.

[10] X. Serra and J. O. Smith, "Spectral modeling synthesis: a sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[11] S. N. Levine, "Audio representations for data compression and compressed domain processing," Ph.D. dissertation, Stanford University, Stanford, Calif, USA, 1998.

[12] H. Purnhagen and N. Meine, "HILN-the MPEG-4 parametric audio coding tools," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '00)*, vol. 3, pp. 201–204, Geneva, Switzerland, May 2000.

[13] C.-P. Wu, P.-C. Su, and C.-C. J. Kuo, "Robust and efficient digital audio watermarking using audio content analysis," in *Proceedings of Security and Watermarking of Multimedia Contents II: Audio Watermarking*, vol. 3971 of *Proceedings of SPIE*, pp. 382–392, San Jose, Calif, USA, January 2000.

[14] M. Ali, "Adaptive signal representation with application in audio coding," Ph.D. dissertation, University of Minnesota, Minneapolis, Minn, USA, 1996.

[15] M. F. Mansour and A. H. Tewfik, "Time-scale invariant audio data embedding," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 10, pp. 993–1000, 2003.

[16] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Systems Journal*, vol. 35, no. 3-4, pp. 313–336, 1996.

[17] X. Dong, M. F. Bocko, and Z. Ignjatovic, "Data hiding via phase manipulation of audio signals," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 5, pp. 377–380, Montreal, QC, Canada, May 2004.

[18] B. Chen and G. 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, 2001.

[19] R. Petrovic, "Audio signal watermarking based on replica modulation," in *Proceedings of the 5th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Service (TELSIKS '01)*, vol. 1, pp. 227–234, Nis, Yugoslavia, September 2001.

[20] S. Shin, O. Kim, J. Kim, and J. Choil, "A robust audio watermarking algorithm using pitch scaling," in *Proceedings of the 14th International Conference on Digital Signal Processing (DSP '02)*, pp. 701–704, Pine Mountain, GA, USA, October 2002.

[21] L. Girin and S. Marchand, "Watermarking of speech signals using the sinusoidal model and frequency modulation of the partials," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 1, pp. 633–636, Montreal, QC, Canada, May 2004.

[22] Y.-W. Liu and J. O. Smith, "Watermarking sinusoidal audio representations by quantization index modulation in multiple frequencies," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 5, pp. 373–376, Montreal, QC, Canada, May 2004.

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

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

[25] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*, Springer, Berlin, Germany, 1990.

[26] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proceedings of the IEEE*, vol. 81, no. 10, pp. 1385–1422, 1993.

[27] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards*, Kluwer Academic Publishers, Boston, Mass, USA, 2003.

[28] I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital Watermarking*, Morgan Kaufmann, San Francisco, Calif, USA, 2002.

[29] E. Terhardt, "Calculating virtual pitch," *Hearing Research*, vol. 1, no. 2, pp. 155–182, 1979.

[30] M. Abe and J. O. Smith, "Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of FFT magnitude peaks," in *Proceedings of the 117th Audio Engineering Society Conventions and Conferences (AES '04)*, p. 6256, San Francisco, Calif, USA, October 2004.

[31] E. G. Shower and R. Biddulph, "Differential pitch sensitivity of the ear," *Journal of the Acoustical Society of America*, vol. 3, no. 1A, pp. 275–287, 1931.

[32] C. C. Wier, W. Jesteadt, and D. M. Green, "Frequency discrimination as a function of frequency and sensation level," *Journal of the Acoustical Society of America*, vol. 61, no. 1, pp. 178–184, 1977.

[33] F.-G. Zeng, Y.-Y. Kong, H. J. Michalewski, and A. Starr, "Perceptual consequences of disrupted auditory nerve activity," *Journal of Neurophysiology*, vol. 93, no. 6, pp. 3050–3063, 2005.

[34] Y.-W. Liu, "Audio watermarking through parametric synthesis models," in *Digital Audio Watermarking Techniques and Technologies: Applications and Benchmarking*, N. Cvejic, Ed., Idea Group, Hershey, Pa, USA, 2007.

[35] L. L. Scharf and L. T. McWhorter, "Geometry of the Cramer-Rao bound," in *Proceedings of the 6th IEEE SP Workshop on Statistical Signal and Array Processing*, vol. 31, no. 3, pp. 301–311, Victoria, BC, Canada, October 1992.

[36] M. Wolters, K. Kjörling, D. Homm, and H. Purnhagen, "A closer look into MPEG-4 high efficiency AAC," in *Proceedings of the 115th Audio Engineering Society Conventions and Conferences (AES '03)*, New York, NY, USA, October 2003.

[37] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[38] P. Kabal, "An examination and interpretation of ITU-R BS.1387: perceptual evaluation of audio quality," Tech. Rep., Department of Electrical & Computer Engineering, McGill University, Montreal, Canada, 2003. http://www-mmsp.ece.mcgill.ca/Documents/Software/.

[39] V. Pless, *Introduction to the Theory of Error-Correcting Codes*, Wiley-Interscience, New York, NY, USA, 3rd edition, 1998.

[40] J. J. Eggers, R. Bäuml, R. Tzschoppe, and B. Girod, "Scalar Costa scheme for information embedding," *IEEE Transactions on Signal Processing*, vol. 51, no. 4, pp. 1003–1019, 2003.

[41] P. Moulin and R. Koetter, "Data-hiding codes," *Proceedings of the IEEE*, vol. 93, no. 12, pp. 2083–2126, 2005.