Hindawi Publishing Corporation
Research Letters in Signal Processing
Volume 2008, Article ID 346767, 5 pages
doi:10.1155/2008/346767

Research Letter

Psychoacoustic Music Analysis Based on the Discrete Wavelet Packet Transform

Xing He[1] and Michael S. Scordilis[2]

[1] *BrainMedia LLC, New York, NY 10016-5902, USA*
[2] *Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33146, USA*

Correspondence should be addressed to Xing He, starchina@gmail.com

Received 14 November 2007; Accepted 24 January 2008

Recommended by Mark Liao

Psychoacoustical computational models are necessary for the perceptual processing of acoustic signals and have contributed significantly in the development of highly efficient audio analysis and coding. In this paper, we present an approach for the psychoacoustic analysis of musical signals based on the discrete wavelet packet transform. The proposed method mimics the multiresolution properties of the human ear closer than other techniques and it includes simultaneous and temporal auditory masking. Experimental results show that this method provides better masking capabilities and it reduces the signal-to-masking ratio substantially more than other approaches, without introducing audible distortion. This model can lead to greater audio compression by permitting further bit rate reduction and more secure watermarking by providing greater signal space for information hiding.

1. INTRODUCTION

The development of high-quality audio compression methods [1, 2] and effective audio watermarking [3, 4] have benefited greatly from the successful integration of psychoacoustic models. Audio compression methods try to represent the original audio with as low bit rate as possible. High audio quality is achieved by rendering quantization noise inaudible. Audio watermarking techniques, on the other hand, hide the information into the host signal by utilizing auditory masking effects which make possible to keep the embedded watermarks inaudible. Short-time Fourier transform (STFT) has typically been used to obtain a time-varying spectral representation of the signal in the derivation of most psychoacoustic models [2, 5, 6]. Due to the fixed length of analysis windows, the STFT can only provide averaged frequency information of the signal and it lacks the flexibility of arbitrary time-frequency localization [2], which is in striking contrast with the unpredictably dynamic spectral-temporal profile of information-carrying audio signals. Wavelet analysis, on the other hand, presents an attractive alternative by providing multiresolution capability. Specifically, its long windows analyze low frequency components and achieve high frequency resolution while progressively shorter windows analyze higher frequency components to achieve better time resolution.

There have been several attempts using wavelet-based psychoacoustic models in audio. In [7], Sinha and Tewfik proposed a wavelet audio coding scheme by first calculating the masking thresholds in the frequency domain via the fast Fourier transform (FFT). Those thresholds were translated into the wavelet domain and used to ensure transparent audio coding by keeping the error caused either by quantization or by the approximation of the wavelet coefficients below the threshold.

In [8], Zurera et al. proposed a method to effectively represent the psychoacoustic model information in the wavelet domain when low-selectivity filters were used to implement the wavelet transform. Masking thresholds were first calculated in the frequency domain by using the FFT. Those thresholds were partitioned by the equivalent filter magnitude frequency response of the corresponding filter bank branch. Assuming orthogonality of subband signals and quantization noise with white noise-like properties in each subband, the overall masking threshold was represented in the wavelet domain and used tohide the quantization noise.

In [9], Carnero and Drygajlo used a frame-synchronized fast wavelet packet transform algorithm to construct a wavelet domain psychoacoustic model representation. Simultaneous masking thresholds were estimated in a manner similar to transform coding proposed by Johnston [5]. Temporal masking was included by considering the energy within each subband. Final masking thresholds were obtained by considering both simultaneous and temporal masking as well as the band thresholds in absolute quiet. Since this model was tailored specifically for speech signals, its effectiveness on wideband music signals is untested.

The above psychoacoustic modeling methods are either computationally expensive [7, 8], have limited time-frequency representation capabilities by relying on the Fourier transform for the computation of the psychoacoustic model, or approximate the critical bands suboptimally [9], which may often result in objectionable audible distortion in the reconstructed signal. In this paper, we present a new psychoacoustic model computed entirely in the wavelet domain. The STFT is avoided by having wavelet analysis results incorporated in effective simultaneous and temporal masking. Furthermore, the proposed model introduces a wavelet packet-based decomposition that better approximates critical bands distribution. The proposed model maintains perceptual transparency and provides an attractive alternative appropriate for audio compression and watermarking.

The rest of the paper is organized as follows: in Section 2, we introduce the enhanced psychoacoustic model based on the discrete wavelet packet transform (DWPT). Experimental evaluation results are shown in Section 3, followed by the conclusion in Section 4.

2. DWPT-BASED PSYCHOACOUSTIC MODEL

While related analysis techniques [7–9] share a similar general structure, the proposed psychoacoustic model achieves an improved decomposition of the signal into 25 critical bands using the discrete wavelet packet transform (DWPT). This results in a spectral partition which approximates the critical band distribution much closer than before. Furthermore, the masking thresholds are computed entirely in the wavelet domain.

2.1. Signal decomposition with the discrete wavelet packet transform (DWPT)

The discrete wavelet packet transform can conveniently decompose the signal into an auditory critical band-like partition [7–9]. In this work, we divided the input audio signal into 25 standard subbands using DWPT in the manner shown in Figure 1, where the band index is enumerated from 1 to 25 to cover the entire audible spectrum.

A signal decomposition into critical bands resulting from wavelet analysis needs to satisfy the spectral resolution requirements of the human auditory system. On the other hand, the selection of the wavelet basis also is critical for meeting the required auditory temporal resolution, which ranges from less than 10 ms at high frequencies to up to 100 ms at low frequencies [6]. Those constraints make the wavelet base of order 8 (length $L = 16$ samples) a good choice, with specific properties as follows.

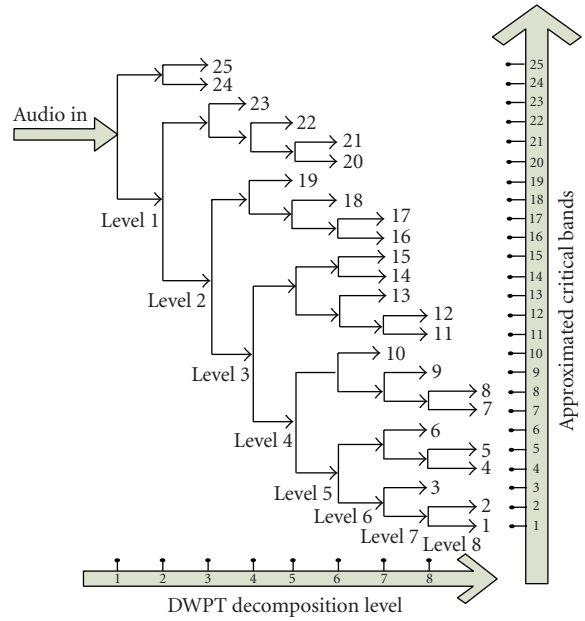

Figure 1: DWPT-based signal decomposition.

The frame length (in samples) at level j ($j = 2, 3, \ldots, 8$) is given by $F_j = 2^j$ and the duration of the analysis window (in samples) at level j is [4, 9]

$$DW_j = (L-1)(F_j - 1) + 1, \tag{1}$$

where L is the length of Daubechies filter coefficients ($L = 16$ in this case). The Daubechies wavelet was selected because it is the most compactly supported wavelet (finer frequency resolution) compared to other wavelet bases with the same number of vanishing moments [10].

For signal bandwidth of 22 kHz, the maximum frame length is 256 samples ($j = 8$) which provides frequency resolution of 22 kHz/256 = 86 Hz. The minimum frame length is 4 samples ($j = 2$) with frequency resolution 22 kHz/4 = 5.5 kHz. The maximum duration of the analysis window is $W_{\max} = 15*(256-1)+1 = 3826$ samples, which at sampling rate of 44.1 kHz corresponds to 87 ms and it applies to the low frequency end, while the minimum duration of the analysis window is $W_{\min} = 15*(4-1)+1 = 46$ samples, or about 1 ms, which applies to the high frequency end.

2.2. Wavelet decomposition evaluation

Wavelet-based approaches to psychoacoustic model implementation are relatively new. In [9], a frame-synchronized fast wavelet packet transform was used to decompose wideband speech into 21 subbands which approximate the critical bands. The spreading function was optimized to speech listening. For wideband audio, [4] has extended that work. 26 critical bands were used and the spreading function was appropriately altered to ensure transparency and inaudibility in audio watermarking applications.

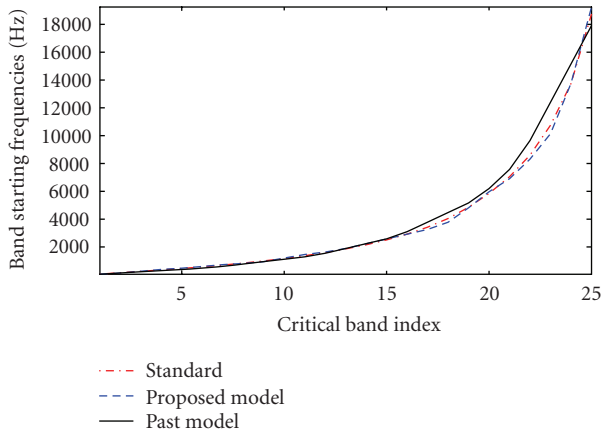

FIGURE 2: Starting frequencies (lower edge) of each critical band.

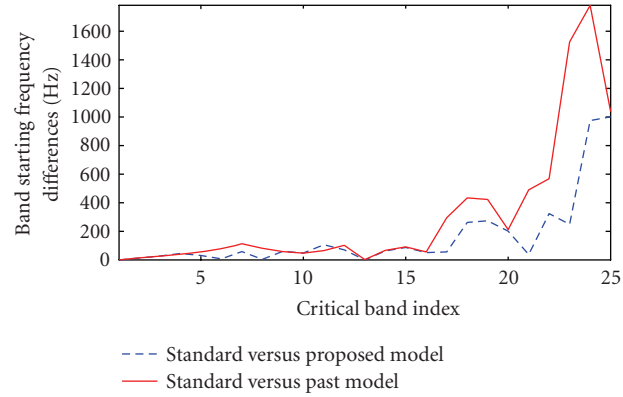

FIGURE 3: Starting frequency differences for each critical band.

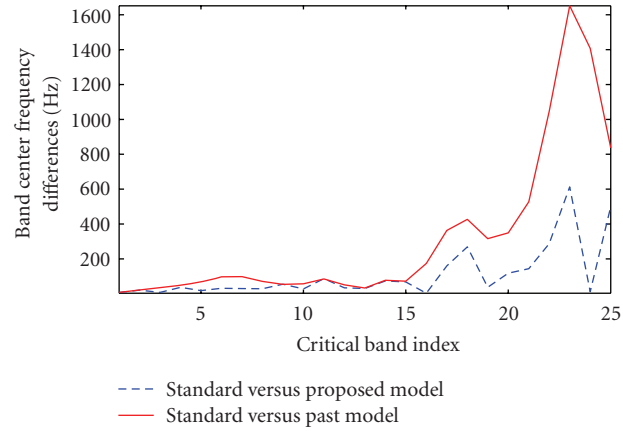

FIGURE 4: Center frequency differences for each critical band.

The critical bands partition obtained by the proposed model is compared to other critical bands approximations introduced and used elsewhere [4, 9]. The degree to which those approaches and the one proposed in this work approximate the standard critical bands partition [5] can be examined by plotting the critical bands starting frequencies, as shown in Figure 2. When the differences in starting frequency are plotted, as shown in Figure 3, it is readily observed that the proposed band partition is substantially closer to the standard, particularly beyond the 16th critical band (frequencies of 2800 Hz and higher). The differences between the two approaches are more striking when critical bands center frequency differences are examined, as depicted on Figure 4, where it can be seen that the proposed approach is considerably closer to the standard. A better approximation to the standard critical bands can provide a more accurate computation of the psychoacoustic model. While this wavelet approach yields a spectral partition that is much closer to the standard critical bands frequencies, the inherent continuous subdivision of the spectrum by a factor of 2 prevents an exact match. However, the overall analysis features of this approach outlined elsewhere in this discussion uphold its overall appeal over competing techniques.

Window size switching similar to [7] is introduced in the proposed psychoacoustic model to mitigate the preecho problem.

Temporal masking is also considered using the method mentioned in [11]. However, compared to other techniques, in this model the entire algorithm operates in the wavelet domain.

3. EXPERIMENTAL PROCEDURES AND RESULTS

The proposed method was evaluated and compared against the standard analysis methods from two useful perspectives: (i) the extent to which portions of the signal power spectrum can be rendered inaudible and therefore removed or altered without any audibly perceived impact, and (ii) the amount of reduction in the sum of signal-to-mask ratio (SSMR) that can be achieved, which is a direct indication that the degree quantization constraints can be relaxed and the coding bit rate can be lowered in signal compression applications, without further loss in perceived quality.

3.1. *Masking of the power spectrum*

The proposed technique was compared against a standard DFT-based approach both quantitatively, in terms of the amount of simultaneous masking provided, as well as qualitatively via listening tests using a variety of musical signals.

In a typical example, the power spectrum of an audio frame (46 ms of audio at 44.1 kHz sampling rate) is depicted in Figure 5, obtained by the square of the magnitude of the DFT coefficients, together with the resulting masking threshold, denoted by the solid line, which was derived according to the perceptual entropy (PE) model used in the MPEG-1 psychoacoustical model 2 [1, 5]. The wavelet power spectrum, or scalogram, was obtained by squaring the magnitude of the wavelet coefficients of the same audio frame and it is shown in Figure 6, together with the associated masking threshold denoted by the solid line, derived by the proposed model.

If the power spectrum length is L and the number of the components below the masking threshold is R (the frequency

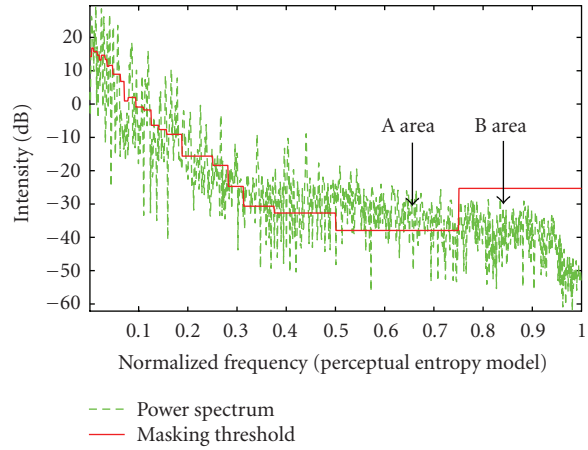

Figure 5: Analysis of a signal frame using the PE model [1, 5].

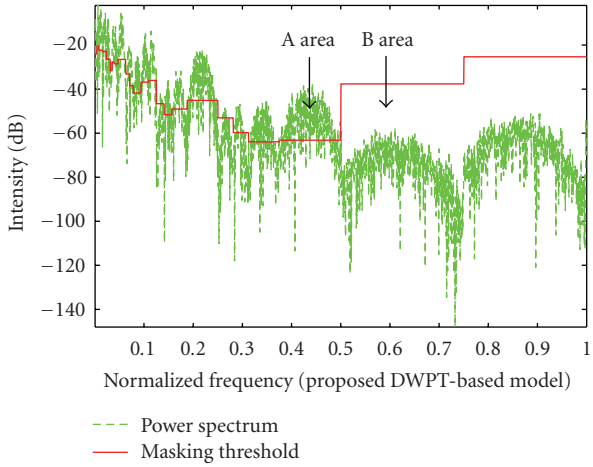

Figure 6: Analysis of a signal frame using the proposed model.

width of B area), then the portion of the removable spectrum is given by

$$Wmc = \frac{R}{L} \times 100\%. \tag{2}$$

From Figures 5 and 6, it can be seen that for this particular example, the DWPT analysis has substantially more spectral components falling under the masking threshold, and are therefore inaudible, than in the Fourier-based analysis.

Five types of musical signals were used in experiments to estimate the percentage of the removable power spectrum for both models. They contained varying musical pieces of CD-quality, which included jazz, classical, pop, country, and rock music. From the results shown in Table 1, it can be seen that an overall gain of about 20% in the extent of masked regions provided by the proposed wavelet method is achieved.

Subjective listening tests were conducted as well, and confirmed that by removing the masked spectral components in each approach and resynthesizing the signal, the processed audio signals are indistinguishable from the original for both the MPEG-based and the proposed technique.

Table 1: Power spectrum portion under masking threshold [5].

Audio type	PE model (%)	Proposed DWPT-based model (%)	Gain (%)
Country	54	73	19
Jazz	50	74	24
Pop	58	74	16
Rock	72	77	5
Classic	44	78	34
AVERAGE	55	75	20

Table 2: Sum of signal-to-masking ratio comparison.

Audio type	PE model (dB)	Proposed DWPT-based model (dB)	Gain (%)
Country	17178	7129	59
Jazz	16108	5447	66
Pop	20061	12156	40
Rock	21266	14411	32
Classic	14756	2075	86
AVERAGE	17874	8244	57

3.2. Signal-to-masking ratio reduction

The ability to facilitate lower bit rates in music compression schemes is another useful consideration in comparing the effectiveness of the two methods. The signal-to-mask ratio (SMR) plays an important role in this process because it is a measure of how high the quantization noise floor can be allowed to raise in the audible signal components. A small SMR indicates that a relatively high noise floor is permissible and therefore fewer bits may be used for coding. The performance metric that captures this effect for the entire spectrum of a particular analysis frame is defined as the average sum of the signal-to-mask ratio (SSMR).

Specifically, if S is the power spectrum of the signal, M is the masking threshold, all in dB and functions of frequency, L is the number of samples in the audio file, then perceptual entropy (PE), which is a lower bound estimate for the perceptual coding of audio signals based on the psychoacoustic model, is defined as [6]

$$\begin{aligned} \text{PE} &= \frac{1}{L}\sum_{i=1}^{L} \max\left\{0, \frac{1}{2}\log_2\left(\frac{10^{(S_i/10)}}{10^{(M_i/10)}}\right)\right\} \\ &= \frac{1}{L}\sum_{i=1}^{L} \max\left\{0, \frac{1}{2}\log_2\left(10^{(\text{SMR}_i/10)}\right)\right\}. \end{aligned} \tag{3}$$

From (3), it can be seen that the reduction of SMR (as long as SMR is positive) will lead to the reduction in PE and consequently a lower bit rate for audio coding by allowing larger quantization noise to be tolerated.

Examining the analysis of the previous example depicted in Figures 5 and 6, areas A and B can be defined in terms of SMR as $A = \{S \mid \text{SMR} \geq 0\}$ and $B = \{S \mid \text{SMR} < 0\}$. In audio compression applications, in area A, which consumes

all allocated bits, the encoded signal must be close enough to the original signal to maintain the quantization noise below the masking threshold.

Let $\mathrm{SMR}_{i,j}$ denote the SMR of the ith sample of area A in the jth signal frame, L_j denote the length of A area in that frame, and G be the number of total frames in the signal analyzed. Then, the sum of SMR (SSMR) in dB for the duration of the signal is given by

$$\mathrm{SSMR} = \sum_{j=1}^{G} \sum_{i=1}^{L_j} \mathrm{SMR}_{i,j}. \tag{4}$$

The two models, proposed and standard, were compared using the same audio material as in the previous test. The results are summarized in Table 2.

As it can be seen from Table 2, in the proposed wavelet-based technique, the SSMR was reduced by as much as 86% (for country music), while the average reduction rate reaches 57%, indicating that a significant decrease in coding bit rate is possible.

4. CONCLUSION

The proposed psychoacoustic model uses the discrete wavelet packet transform to provide multiresolution analysis that closely mimics auditory processing and it is superior to Fourier transform-based techniques both from the computational as well as the resolution perspectives. The auditory critical bands distribution is implemented more accurately than in previous techniques. The model includes simultaneous and temporal masking effects, all computed in the wavelet domain. Experiments conducted on a variety of music signals demonstrate that the proposed method provides broader masking capabilities thus revealing that larger signal regions are in fact inaudible and therefore removable without noticeable effect, a fact that was confirmed in listening tests. The masked regions may be ignored in audio compression thus resulting in lower information rates, or may be used for hiding more information in audio watermarking. Furthermore, the signal-to-masking ratio is further reduced indicating that in coding applications, this approach can lead to further bit rate reduction without quality degradation.

APPENDIX

SHORT INTRODUCTION TO CWT, DWT, AND DWPT

The continuous wavelet transform, CWT, of signal $s(t)$ is defined as

$$\mathrm{CWT}(\alpha, \tau) = \frac{1}{\sqrt{\alpha}} \int s(t) \psi^* \left(\frac{(t-\tau)}{\alpha} \right) dt, \tag{A.1}$$

[10] where * is the complex conjugate operation, t is time, τ is the temporal translation parameter, α is the scaling parameter, and $\psi(t)$ is the transforming function, called mother wavelet. Parameter τ provides the time location of the analysis window, and it varies as the window is shifted through the signal, while α controls the amount of stretching or compressing of the mother wavelet $\psi(\tau)$, which controls the shape of the wavelet.

In discrete time, signal $s(n)$ can be equivalently transformed by the discrete wavelet transform (DWT), which is discrete both in the time and the wavelet domains, and it is defined as [10]

$$\mathrm{DWT}(m, n) = 2^{(-m/2)} \sum_{k} s(k) \psi^* \left(2^{-m} k - n\right). \tag{A.2}$$

This is the discrete version of (A.1), with $\tau = 2^m n$ and $\alpha = 2^m$, where m, n, and k are integers.

The DWT is often implemented by a group of filter banks consisting of half-band high ($\pi/2$ to π) and low pass filters (0 to $\pi/2$). The signal is first divided into high- and low-frequency parts by the high- and low-pass filters, respectively, and the low frequency part is further decomposed into high- and low-frequency parts. The process continues on the low-frequency part until the desired decomposition is achieved. If both the high- and low-frequency parts are recursively decomposed, the DWT turns into the discrete wavelet packet transform (DWPT), which is a more flexible computational structure and it can be incorporated in audio analysis to closely approximate auditory critical bands.

REFERENCES

[1] ISO/IEC 11172-3, "Information technology—coding of moving picture and associated audio for digital storage media at up to about 1.5 Mbits—part 3: audio," 1993.

[2] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–512, 2000.

[3] 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.

[4] Q. Liu, "Digital audio watermarking utilizing discrete wavelet packet transform," M.S. thesis, Institute of Networking and Communication, Chaoyang University of Technology, Taichung, Taiwan, 2004.

[5] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, 1988.

[6] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards*, Kluwer Academic Publishers, New York, NY, USA, 2003.

[7] D. Sinha and A. H. Tewfik, "Low bit rate transparent audio compression using adapted wavelets," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3463–3479, 1993.

[8] M. R. Zurera, F. L. Ferreras, M. P. J. Amores, S. M. Bascón, and N. R. Reyes, "A new algorithm for translating psycho-acoustic information to the wavelet domain," *Signal Processing*, vol. 81, no. 3, pp. 519–531, 2001.

[9] B. Carnero and A. Drygajlo, "Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms," *IEEE Transactions on Signal Processing*, vol. 47, no. 6, pp. 1622–1635, 1999.

[10] I. Daubechies, *Ten Lectures on Wavelets*, vol. 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, SIAM, Philadelphia, Pa, USA, 1992.

[11] B. Lincoln, "An experimental high fidelity perceptual audio coder," *Project in MUS420 Win97*, March 1998.