Low Bit Rate Transparent Audio Compression using Adapted Wavelets

Deepen Sinha, *Student Member, IEEE* and Ahmed H. Tewfik, *Senior Member, IEEE*

Abstract—**This paper describes a novel wavelet based audio synthesis and coding method. The method uses optimal adaptive wavelet selection and wavelet coefficients quantization procedures together with a dynamic dictionary approach. The adaptive wavelet transform selection and transform coefficient bit allocation procedures are designed to take advantage of the masking effect in human hearing. They minimize the number of bits required to represent each frame of audio material at a fixed distortion level. The dynamic dictionary greatly reduces statistical redundancies in the audio source. Experiments indicate that the proposed adaptive wavelet selection procedure by itself can achieve almost transparent coding of monophonic compact disk (CD) quality signals (sampled at 44.1 kHz) at bit rates of 64–70 kilobits per second (kb/s). The combined adaptive wavelet selection and dynamic dictionary coding procedures achieve almost transparent coding of monophonic CD quality signals at bit rates of 48–66 kb/s.**

I. Introduction

REDUCING the bit rate requirement is an important task in the design of audio systems. In many applications, such as the design of multimedia workstations and high quality audio transmission and storage the goal is to achieve transparent coding of hi-fidelity audio signals at the lowest possible bit rates. Typically, the quality of compact disk (CD) signals is used as the standard for high fidelity. These signals are characterized by a wide bandwidth (sampling rate of 44.1 kHz) and a high resolution quantization of each sample of the signal (16 bits/sample pulse code modulation (PCM) quantization scheme) resulting in a high bit rate of 705 kb/s.

Several subband coding [32], [33], [39], [40] and transform coding [4], [20], [21] approaches have been proposed for reducing the bit rate requirement in the above mentioned applications. Some of these methods claim to achieve perceptually transparent coding of monophonic CD quality signals at approximately 96 kb/s. Further reduction in bit rate requirement is an attractive proposition in applications like remote broadcast lines, studio links, satellite transmission of high quality audio, multimedia products, etc.

Manuscript received September 10, 1992; revised May 10, 1993. The guest editor coordinating the review of this paper and approving it for publication was Dr. Takao Nishitani. This work was supported in part by AFOSR under Grant AF/F4 9620-92-7-0134 and in part by a Grant from Texas Instruments.

The authors are with the Department of Electrical Engineering, University of Minnesota, Minneapolis, MN 55455.

IEEE Log Number 9212178.

In other applications the objective is to synthesize music signals at or below 10 kb/s. The transform and subband coders have generally been found to be unsuitable at these rates. On the other hand, several successful low bit rate speech coders have been described in the literature. These coders are almost invariably based on the well known excitation-modulation model of human speech production. The modulation filter is typically an all pole filter which is identified by a linear prediction (LP) analysis over a short frame of speech data. The excitation takes different forms. Since the LP model relies on the human voice production mechanism, it may not be suitable for coding music and other non-speech sounds. However, some recent studies suggest that a variant of this model (the multipulse LPC algorithm) is viable for this task [31].

To achieve good quality at low bit rates an audio compression scheme must exploit the two sources of irrelevancies and redundancies in audio signals: the masking characteristics of the human hearing process and the statistical redundancies in the signal. Most currently known methods for perceptually transparent coding of CD quality signals concentrate on exploiting the masking effect. On the other hand, very low bit rate methods focus on the elimination of statistical redundancies. In this paper, we present a novel approach which employs an *optimal* wavelet based coding method to exploit perceptual masking and a dynamic dictionary based coding method to obtain further bit rate reduction by eliminating source redundancies. The combined procedure results in a very high level of compression.

The adaptive wavelet coding step optimally selects the discrete wavelet transform (DWT) representation of each frame. It also allocates bits to the transform coefficients in an adaptive fashion. The wavelet transform selection and bit allocation procedures are designed to minimize the number of bits required to represent the given frame at a fixed distortion level. We have chosen to implement an adaptive DWT signal representation because the DWT is a highly flexible family of signal representations that may be matched to a given signal [37]. It also provides a good approximation to the Karhunen-Loeve transformation (KLT) of a wide class of stationary and nonstationary processes [36], [41]. Therefore, the DWT is readily applicable to the task of audio data compression. The ability of the adaptive wavelet coding step to exploit the masking properties of human hearing is demonstrated by the fact that almost perceptually transparent encoding of a wide

variety of audio signals was maintained at bit rates of 64–70 kb/s by using this procedure only. Used together, the adaptive wavelet coding and the dynamic dictionary encoding procedures have achieved almost perceptually transparent coding of a variety of audio signals at bit rates of 48–66 kb/s.

The technique that we propose currently requires a long coding delay. Decoding on the other hand can be accomplished in real time. This is not an issue in many applications that we target such as audio synthesis and playback of recorded audio material. Further advances in the understanding of the proposed approach should lead to lower coding delay and computational complexity. It is also worth noting that preliminary studies suggest that our technique may be a viable approach for the low bit rate synthesis of music signals at about 10 kb/s.

The organization of this paper is as follows. In Section II, we introduce the discrete wavelet transform. We discuss the conditions under which the wavelet has a high degree of regularity and methods for constructing such wavelets. In Section III, we review the essential facts about perceptual masking and the masking model used in this paper. Then, we translate the masking threshold condition to the wavelet transform domain. In Section IV, we formulate the problem of choosing the DWT representation that can approximate the audio signal at a given distortion level using a minimum number of bits. We then proceed to present a set of results that simplify this optimization problem. Several issues in the optimization problem discussed above are best answered by a set of suitably designed experiments. This is the focus of Section V. The ingredients of the dynamic dictionary based encoding are discussed in Section VI. In Section VII, we discuss several implementation issues. Finally, we present in Section VIII experimental results to demonstrate the viability of the proposed method. This is followed by a discussion of possible improvements and future research topics.

II. Discrete Wavelet Transformation

Wavelets are a new family of basis functions [8], [23], [24] for the space of square integrable signals. A signal $f(t) \in L^2(\boldsymbol{R})$ can be represented in terms of the translates and dilates of a single wavelet $W(t)$ as

$$f(t) = \sum_{j,m} \sqrt{2^j}\, b(j; m)\, W(2^j t - m) \tag{1}$$

or equivalently,

$$f(t) = \sum_{j=L}^{\infty} \sum_{m=-\infty}^{\infty} \sqrt{2^j}\, b(j; m)\, W(2^j t - m) + \sum_{m=-\infty}^{\infty} \sqrt{2^L}\, a(L; m)\, g(2^L t - m) \qquad \forall L \in \boldsymbol{Z} \tag{2}$$

where

$$a(j; m) = \int_{-\infty}^{\infty} f(t) \sqrt{2^j}\, g(2^j t - m)\, dt$$

$$b(j; m) = \int_{-\infty}^{\infty} f(t) \sqrt{2^j}\, W(2^j t - m)\, dt.$$

Such an expansion provides a multiresolution analysis of $f(t)$. Specifically, let ω_0 be the center frequency of the band-pass function $W(t)$. The coefficients $b(j; m)$ carry information about $f(t)$ near the frequency $2^j\omega_0$ and the time instant $2^{-j}m$. These coefficients are called the *detail* coefficients. The second summation in (2) provides a low pass filtered version of $f(t)$ up to scale 2^L. The coefficients $a(j; m)$ are therefore referred to as *approximation* coefficients at scale 2^j.

The wavelet $W(t)$ must satisfy certain conditions that ensure that the above expansion holds for any square integrable function. Specifically, $W(t)$ is obtained from a *scaling function* $g(t)$ as

$$W(t) = \sum_{k=0}^{K-1} (-1)^k c_{1-k}\, g(2t - k). \tag{3}$$

The scaling function in turn obeys a dilation equation

$$g(t) = \sum_k c_k\, g(2t - k). \tag{4}$$

Here, we shall restrict our attention to compact support wavelets [8]. For these wavelets the coefficients c_k defining $g(t)$ can be nonzero only for $0 \le k \le K - 1$. The coefficients c_k satisfy a number of constraints.

The first condition on the coefficients c_k arises from the fact that $g(t)$ is a low-pass function normalized such that $\int g(t)\, dt = 1$. From this we conclude

$$\sum_{k=0}^{K-1} c_k = 2. \tag{5}$$

To ensure that the translates of $g(t)$ are orthogonal we require that

$$\sum_{k=0}^{K-1} c_k c_{k+2m} = 2\delta_{0,m} \tag{6}$$

where $\delta_{0,m}$ is the Kronecker delta function. Let $P(e^{j\omega}) = \frac{1}{2} \sum_k c_k e^{jk\omega}$. Condition (6) above is equivalent to the "power complimentary condition"

$$|P(\omega)|^2 + |P(\omega + \pi))|^2 = 1. \tag{7}$$

The above construction of the wavelet $W(t)$ together with (6) and additional technical conditions [8] ensures that all translates and dilates of $W(t)$ are orthogonal and that $W(2^j t - k)$ is orthogonal to $g(2^l t - m)$ for all $j \ge l$ and all m and k. It also implies that span $\{\sqrt{2^j}\, W(2^j t - m)\}_{j=-\infty}^{J-1}$ = span $\{\sqrt{2^J}\, g(2^J t - m)\}$.

Conditions (5) and (6) are not enough to construct *useful* wavelets. The reason is that they can lead to wavelet decomposition with not enough regularity, i.e., the coefficients $f_{j,k}$ may decay too slowly to zero. Regularity of the decomposition is ensured by requiring that $W(t)$ has a

large number of vanishing moments. It may be shown that imposing the additional requirement that

$$\sum_k (-1)^k k^m c_k = 0 \qquad m = 0, 1, 2, \cdots, p-1 \quad (8)$$

leads to a wavelet $W(t)$ with p vanishing moments. The moment condition (8) is equivalent to requiring that the polynomial $P(\omega) = \frac{1}{2}\Sigma_k c_k e^{ik\omega}$ has a zero of order p at $\omega = \pi$. It will be seen in the following sections that wavelets with large number of vanishing moments are specially attractive for our audio compression method. We will return to this condition shortly and discuss a method for constructing such wavelets.

In practice, the wavelet transform coefficients, i.e., $a(j; m)$ and $b(j; m)$ and computed recursively from $a(j+1; m)$ using an efficient *pyramid* [23] algorithm. In particular, it is not necessary to explicitly compute the shape of $g(t)$ and $W(t)$. It may be shown that if J is large then $a(J, m) \approx f(m/2^J)$. Since we usually deal with a finite set of N data samples, we take $2^J = N$ to be the finest scale. At this scale we let $a(J; m)$ be equal to the samples of $f(t)$. Furthermore, we assume that the given data is periodic with period N. With this assumption, it may be shown that the DWT acts as an orthonormal linear transform T: $\boldsymbol{R}^N \mapsto \boldsymbol{R}^N$ [34]. Its matrix representation $\boldsymbol{Q}$ is fully determined by the coefficients c_k. The matrix $\boldsymbol{Q}$ corresponds to passing the (periodized) data through a cascade of banks of perfect reconstruction low pass and high pass filters $P(\omega)$ and $H(\omega)$ respectively followed by decimators. In particular, the filters are arranged in a tree structure as shown in Fig. 1. The leaf nodes in this tree correspond to *subbands* of the wavelet decomposition. We note that $\boldsymbol{Q}$ consists of circular blocks (one block for each subband), of size $M \times N$, where $M = (N/2^m)$ for a band at depth m in the decomposition tree. Within each block each row is a circularly shifted version of previous row by 2^m samples.

It is shown in [8] that the wavelet $W(t)$ corresponding to a K coefficient sequence $\{c_k\}$ will have a support of length $K-1$. Furthermore, such a wavelet can have at most $K/2$ vanishing moments. It has also been shown in [28] and [42] that all valid sequences $\{c_k\}_{k=0}^{k=K-1}$ are parameterized by $(K/2)-1$ "angle" parameters. Each of these parameters can take values in the range $[0, 2\pi]$. This parametrization is quite important in solving the optimization problems described in the rest of this paper.

In this paper we will concentrate on wavelets of a given finite support and a maximal number of vanishing moments. Proposition 4.5 in [8, p. 977] provides a recipe for constructing all K coefficient sequences $\{c_k\}$ that correspond to wavelets with a maximal number of $K/2$ vanishing moments. Note that all K coefficient sequences $\{c_k\}$ that correspond to wavelets with a support of length $K-1$ and a maximal number of $K/2$ vanishing moments have discrete time Fourier transforms (DTFT) with the same magnitude. Hence, the low pass filters $P(\omega) \equiv \frac{1}{2}\Sigma_k c_k e^{j\omega k}$ that correspond to all such sequences will have identical magnitude responses. The high pass filters $H(\omega)$ shown in Fig. 1 that correspond to these sequences will also have identical magnitude responses.

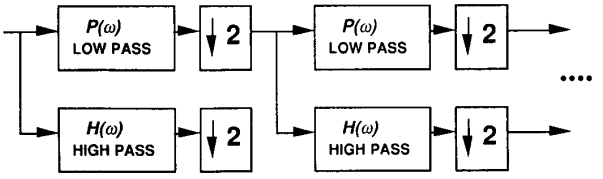
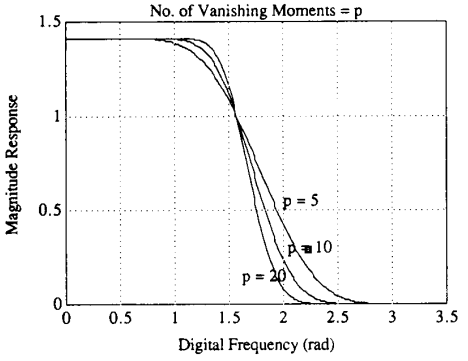
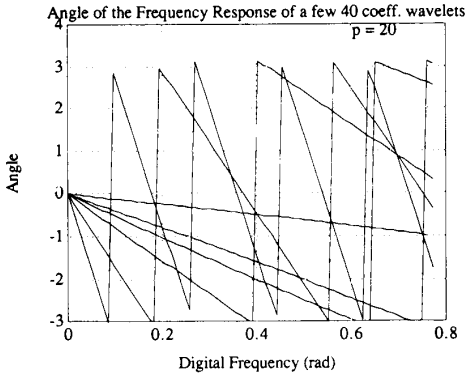


Fig. 1. Wavelet decomposition tree.



(a)



(b)

Fig. 2. Frequency response of maximum vanishing moment filters: (a) Magnitude response of 5, 10, and 20 moment filters; (b) Phase response (wrapped) of a few 20 moment wavelets.

It may also be noted that if the $2p$ tap filter $P(\omega)$ corresponding to the impulse response $\{c_k/2\}$ has maximum number of p zeros at $\omega = \pi$, it behaves almost like a true low-pass filter. For large values of p the magnitude response $|P(\omega)|$ is nearly constant over the pass band. The transition band of $P(\omega)$ is also very narrow. This is illustrated in Fig. 2(a), where the magnitude responses $|P(\omega)|$ of all such $2p$ tap filters with p zeros at $\omega = \pi$ is shown with $p = 5$, 10, and 20. Note also that even though the magnitude response of all such $2p$ tap filters with p zeros at $\omega = \pi$ is identical, the filters do differ in terms of their phase responses. In particular, the filters $P(\omega)$ are nearly linear phase filters and their phase responses may be approximated, to a first degree, by a delay. An illustration

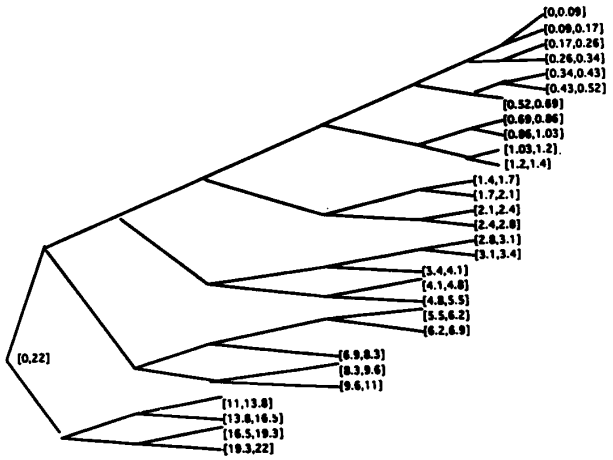


Fig. 3. Wave Packet decomposition tree used in the audio coding application. The numbers in the figure refer to the lower and higher cutoff frequencies in kHz of each of the 29 critical bands that we use.

of this is in Fig. 2(b) where, the phase response of a few 40 tap filters (with 20 vanishing moments) is shown.

Using [8, Proposition 4.5, p. 977] we have found that for $N = 10$, 20, 30, 40, and 60 there are respectively 4, 32, 1024, and 32 768 wavelets with the maximum possible vanishing moments. We have also generated all the sequences $\{c_k\}$ that correspond to these wavelets. These sequences play an important role in the optimization problem discussed in Section IV.

Wave Packet Representation and M-Band Wavelet Transforms

Wave packet representations have been proposed as an extension of the wavelet transformation. In a usual wavelet transform only the approximation at a given scale is further decomposed. In terms of the filtering scheme shown in Fig. 1, only the output of the low pass filter $P(\omega)$ is further decomposed. In contrast, in a wave packet decomposition [7], the 2-band wavelet filter banks of Fig. 1 are used to split both the low pass and high pass bands (as opposed to only the decomposition of low frequency bands as in Fig. 1). This type of decomposition is represented by a binary tree in which one has the freedom to stop or continue the decomposition at any node. Several choices for a basis are thus possible. In [7] an entropy criterion is used to match the decomposition tree to a given data set. While the entropy criterion is not pertinent from our point of view, the binary tree decomposition is attractive for the audio coding application because of the following two reasons. First, if a wavelet with a large number of vanishing moments is used, a precise specification of the pass bands of each subband in the wavelet decomposition is possible. These pass-bands are nonoverlapping. Secondly, psycho acoustic studies of the human ear suggest that a frequency to "bark" transformation needs to be performed to accurately model the frequency dependent sensitivity of human ears to quantization errors [30]. Such a transformation is accomplished in audio coding systems by dividing the frequency range into critical bands. Quantization noise power is integrated over these bands by the human ear. Hence, coding errors for signal components in a critical band have equal weight and these components are typically grouped together during bit allocation. The critical bands that correspond to a sampling rate of 44.1 kHz are given in [20]. We use a wavelet tree structure which closely mimics this critical band division. It is shown in Fig. 3. The lower and higher cutoff frequencies of each of the 29 critical bands are also listed in kHz units on this figure.

Note, that precise specification of the pass-bands can also be accomplished by using an M-band wavelet transforms [12], [16], [43]. In such an approach, one uses an appropriately designed perfect reconstruction filter bank that consists of M filters to decompose the signal at each stage.

III. Auditory Masking

Auditory masking is a phenomenon whereby a weak (noise) signal is made inaudible by a simultaneously occurring stronger (audio) signal. Because of masking perceptually transparent quality in a coding system can be maintained as long as the power spectrum of the reconstruction errors is below a signal dependent threshold at each frequency. Computation of this threshold is based on several considerations. It has been found that masking depends on the frequency and characteristics (tone or noise) of both the masking and masked signal. In general, sound power at one frequency masks the noise power at that frequency and nearby frequencies. Masking of a tonal signal by another tone or by a narrow band noise has been studied in detail [19], [22].

In our coding method we use a model for masking proposed in [39]. In this model the shape of the masking threshold of a pure tone is approximated by

$$T(f_m, f) = \begin{cases} T_{\max}(f_m)\left(\dfrac{f}{f_m}\right)^{28}, & f \le f_m \\ T_{\max}(f_m)\left(\dfrac{f}{f_m}\right)^{-10}, & f > f_m. \end{cases} \tag{9}$$

In the above expression f and f_m are respectively the frequencies of the masking and the masked signals. $T_{\max}(f_m)$ is the relative masking threshold at the masking signal frequency [19]. Fig. 4 provides a plot of the relative masking thresholds $T_{\max}(f_m)$ that we have used in our work. The relative masking thresholds depend on the frequency and tonality (i.e., tone or noise like nature) of the signal. We have estimated values for $T_{\max}(f_m)$ based on the results of [20], [22]. Specifically, we took the relative threshold to be constant over each critical band. The results of [20] suggest that for tone like maskers the relative threshold should be set to $-(14.5 + i)$ dB, where i is the critical band index (i.e., bark frequency) [20], [30]. On the other hand, for noise like maskers the relative threshold may be taken independent of the bark frequency to be

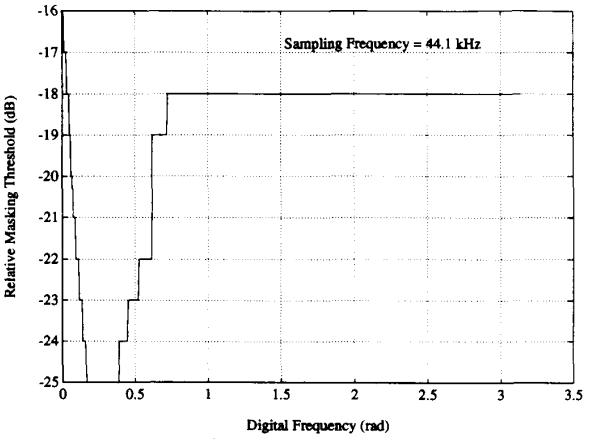


Fig. 4. Relative threshold $T_{\max}(f_m)$, used in this paper.

-5.5 dB. An accurate estimate of the relative threshold, therefore, will require estimating the tonality of the signal component in a critical band [20]. For simplicity, we have used in our present work a composite value for the relative threshold $T_{\max}(f_m)$ based on the idea that a signal in a lower critical band is more tone like in nature while a signal in a higher critical band is more noise-like. Also, note that the higher critical bands are also wider. Therefore, we have set the relative threshold to approximately $-(14.5 + i)$ dB in the lowest frequency range of 0–2.5 kHz. We then raised the relative threshold gradually at frequencies above 2.5 kHz using the results of [22]. The threshold was not raised all the way up to -5.5 dB. Instead, it was frozen at a value of about -18 dB. This more conservative estimate of $T_{\max}(f_m)$ was used at higher frequencies because we do not compute an accurate estimate of tonality.

Given an audio segment we compute the threshold of masking as follows: We begin by estimating the power spectrum $S(f)$ of the segment using a fast Fourier transform (FFT). The masked power at a frequency f due to a signal component at frequency f_m is found by multiplying $S(f_m)$ with $T(f_m, f)$. We assume that masking is additive. Therefore, an estimate of the total masked power at each frequency f is computed by adding the masked power due to the components of the signal at each frequency. We took the perceptual threshold to be constant over each wavelet (critical) band. The value of this constant is equal to the minimum of the estimated masked powers at all frequencies within that band. The estimate of the perceptual threshold is now compared with the absolute minimum threshold (or the threshold of hearing) [20] to ensure that we never demand an accuracy higher than the highest sensitivity of the human ear. This is then the final estimate for the masked noise power $S_n(f)$. A listener will tolerate an additive noise in the reproduced audio signal as long as the power specturm of the noise power is less than $S_n(f)$ at each frequency f.

It may be noted that the simple masking model used here has several drawbacks. First, the shape of $T(f_m, f)$ as given in (11) is valid for masking by tonal signals. Specifically, the interfrequency masking effects are better approximated by (11) for tone like masking signals. However, to simplify the computation of the masking threshold we have used the same functional shape of $T(f_m, f)$ for all types of signals even though the estimates of interfrequency masking will not be accurate when the signal is more noise-like. As mentioned above, we conservatively adjust the relative masking threshold $T_{\max}(f_m)$ to account for masking by noise-like signals. Adjusting the shape of $T(f_m, f)$ as well should yield better results. Second, the model is based on psycho acoustic studies for the masking of a single tone like signal [15]. The reconstruction error may not be masked if it contains several, possibly nontonal, components. This is recognized in [20], where the perceptual threshold is renormalized, and in [38] where a constrained minimization problem is solved to account for multiple noise sources. Finally, masking is assumed to be additive. Experiments in [17] suggest that a power law rule of addition should be used instead. The appropriate power law was found to be signal dependent. Despite of these drawbacks the simple masking model that we have described is attractive from an implementation point of view and yields reasonable coding gains when used in conjunction with the coder that we discuss in the rest of this paper.

Note that masking is both a frequency domain and time domain phenomenon. The frequency domain results that we described above provide a vehicle for exploiting perceptual masking in a practical coder. However, these results are based on the assumption that both the masking signal and the masked do not change with time. It is therefore necessary to ensure that time domain masking constraints are also satisfied when sudden changes in the signal characteristics occur. We will come back to this issue in Section VII.

Masking Constraint in the Wavelet Domain

We now explain how the perceptual masking threshold can be incorporated within the framework of a wavelet transform based coder. Denote by $\boldsymbol{e}$ the $N \times 1$ error vector consisting of the values of the discrete Fourier transform of the error in reconstructing the signal from a sequence of approximate wavelet coefficients. Furthermore, let $\boldsymbol{R_D}$ be a diagonal matrix with entries equal to discretized values of $1/S_n(f)$. The above discussion implies that the reconstruction error due to quantization or approximation of the wavelet coefficients corresponding to the given audio signal may be made inaudible as long as

$$|e_i|^2 r_{ii} \leq 1, \qquad \text{for } i = 1, \cdots, N \tag{10}$$

where e_i is the ith component of $\boldsymbol{e}$ and r_{ii} is the ith diagonal entry of $\boldsymbol{R_D}$. Since we can always fit an ellipsoid inside a multidimensional rectangle, a sufficient condition to ensure that (10) will be satisfied is given by

$$\boldsymbol{e}' \boldsymbol{R_D} \boldsymbol{e} \leq 1. \tag{11}$$

In (11) e' denotes the complex conjugate transpose of vector e. Equation (11) above puts the perceptual norm criterion into a structure that is very suitable for the wavelet transform encoder. However, it is overly restrictive and leads to artificially high estimates for the bit rate requirement. To overcome this, we relax the upper bound in (11) to N. The new perceptual norm criterion then is

$$e'R_D e \leq N. \tag{12}$$

We note that (12) will be sufficient to ensure that the conditions in (10) are satisfied as long as the coding scheme that we use equalizes or "whitens" the weighted frequency domain noise powers $e_i^2 r_{ii}$. The results that we will present in Section IV and experimental evidence show that this is indeed the case for the coder that we propose here.

We now translate (12) into the wavelet transform domain. Note that (12) is equivalent to

$$e_q' QW' R_D WQ' e_q \leq N \tag{13}$$

where e_q is the $N \times 1$ vector consisting of values of the error in the quantization of wavelet coefficients. In (13) Q and W are respectively the wavelet transform and the discrete Fourier transformation matrices. Q' and W' denote respectively the complex conjugate transpose of Q and W.

IV. Wavelet Optimization Approach to Audio Data Compression

We use the following approach to compress an audio signal. The audio data is first divided into overlapping analysis frames. The segmentation procedure that we use to divide the signal is described in Section VII. For each analysis frame an optimum wavelet representation is then selected to minimize the number of bits required to represent the frame while keeping any distortion inaudible. The wavelet selection step involves choosing an analysis wavelet and allocating bits to each coefficient in the resulting wavelet representation. The success of this scheme is based on the fact that only a fraction of the wavelet coefficients corresponding to a frame are allocated a nonzero number of bits and that these coefficients may be encoded using a small number of bits as we argue below.

Fig. 5 shows a signal vector represented by a particular choice of a basis. The radius of the sphere shown in Fig. 5(a) is equal to the norm of the time domain signal. Also shown in Fig. 5 is the error ellipse corresponding to the perceptual seminorm discussed in Section III. Note, that the audio segment may be represented using any vector whose tip lies inside the error ellipse with no perceptual distortion. Hence, the projection of the error ellipsoid along each coordinate axis specifies the coarsest quantization that can be used along that axis without producing any perceptual degradation. Therefore, a large projection along a particular coordinate axis implies that we need to use only a small number of bits to quantize that coordinate. Exploiting this fact, we can obtain a low bit rate

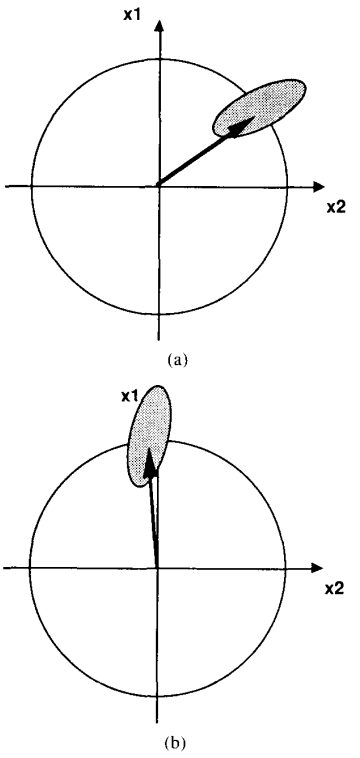


Fig. 5. Illustration of audio compression by optimal basis selection: (a) Signal and error ellipse; (b) Signal and error ellipse under a new choice of the basis.

representation of the signal as follows: We identify a rotation of the vector representation of the signal via a unitary wavelet transformation, as illustrated in Fig. 5(b), to achieve two desirable results. First, we make the projection of the signal vector along most coordinate directions about the same as that of the error ellipsoid. The signal vector projections along these coordinate directions can; therefore, either be neglected and set to zero, or encoded using a small number of bits without producing any perceptual degradation. Second, we make the projection of the error ellipsoid large along the remaining coordinate directions. The signal vector projections along these directions can then be encoded using a small number of bits. Since the wavelet transform is a family of orthogonal basis it provides the flexibility of choosing the unitary transform that best achieves these two desirable results.

Optimization Criterion

Let us denote by x a given audio frame. Furthermore, let $x^q = Qx$ denote its wavelet transform. Finally, let Θ the be the vector of parameters that completely specifies all wavelets of a given support size or equivalently all sequences $\{c_k\}$ of a given length (c.f., Section II and [28], [42]). Our objective is to minimize the number of bits required to represent the frame using a specific wavelet subject to keeping any distortion inaudible. Here, we shall use a (suboptimal) scalar adaptive quantization technique

for the transform coefficients. Further reduction in bit rate can be achieved by jointly quantizing groups of these coefficients.

Specifically, let $R_k(\Theta)$ be the number of bits assigned to the quantization of the kth transform coefficient $x_k^q(\Theta)$ when the wavelet identified by the vector Θ is used to decompose frame $\boldsymbol{x}$. Our goal is to minimize

$$\boldsymbol{R}(\Theta) = \sum_{k=1}^{N} R_k(\Theta) \tag{14}$$

by properly choosing Θ and the number of bits $R_k(\Theta)$ assigned to the quantization of each transform coefficient $x_k^q(\Theta)$. The minimization must be done under the constraint (13) on the perceptual seminorm of the approximation or encoding error. This constraint may be rewritten as

$$\sum_{i,j} e_i(\Theta)\, w_{ij}(\Theta)\, e_j(\Theta) \leq N \tag{15}$$

where $e_i(\Theta)$ is the error in encoding the ith wavelet coefficient $x_i^q(\theta)$, and

$$w_{ij}(\Theta) = \begin{cases} \rho_{ij}(\Theta) & \text{if } i = j \\ 2\Re\{\rho_{ij}(\Theta)\} & \text{if } i \neq j \end{cases} \tag{16}$$

where $\rho_{ij}(\Theta)$ is the (i, j)th element of the matrix $\boldsymbol{Q}(\Theta)\boldsymbol{W}'\boldsymbol{R_D}\boldsymbol{W}\boldsymbol{Q}(\Theta)'$ (13) and $\Re\{\rho_{ij}(\Theta)\}$ denotes the real part of $\rho_{ij}(\Theta)$. Note that the diagonal elements $\rho_{ii}(\Theta)$ of $\boldsymbol{Q}(\Theta)\boldsymbol{W}'\boldsymbol{R_D}\boldsymbol{W}\boldsymbol{Q}(\Theta)'$ are nonnegative constants (see Appendix). Hence, the weights $w_{ii}(\Theta)$ will also be nonnegative. We can simplify (15) as follows:

First, note that the representation levels used by any reasonable adaptive quantizer for a scalar a are directly proportional to the dynamic range of a. Using this fact and known exact and approximate expressions for the quantization errors in various types of quantizers [18], [10], we can assume that the error $e_k(\Theta)$ that results from any adaptive quantization of the kth wavelet coefficient $x_k^q(\Theta)$ is upper bounded by

$$e_k^2(\Theta) \leq \epsilon (x_k^q(\Theta))^2 2^{-2R_k(\Theta)} \tag{17}$$

where ϵ is a constant that depends on the quantizer that is used. Thus, (15) will be satisfied if we can guarantee that

$$\epsilon \sum_{i,j} |w_{ij}(\Theta)| \sqrt{(x_i^q(\Theta))^2 2^{-2R_i(\Theta)} (x_j^q(\Theta))^2 2^{-2R_j(\Theta)}} \leq N. \tag{18}$$

We can further simplify the above constraint by neglecting all cross-terms in (18) for which $i \neq j$. We can justify this simplification as follows:

First, we can show that the matrix $\boldsymbol{Q}(\Theta)\boldsymbol{W}'\boldsymbol{R_D}\boldsymbol{W}\boldsymbol{Q}(\Theta)'$ is nearly diagonal when Θ corresponds to a wavelet with a large number of vanishing moments (see Proposition A.1 in Appendix). While we do not know *a priori* that the optimal wavelet for encoding the given audio segment will have a large number of vanishing moments, our experiments suggest (see below) that there is only a small increase in bit rate estimates if the optimization is limited to the set of wavelets of support length $K - 1$ and a maximal number $K/2$ of vanishing moments as opposed to all possible wavelets with a support of length $K - 1$. This is specially true when K is large. We note also that the solution to the approximate problem that results from neglecting the cross-terms in (18) does indeed provide highly compact wavelet representations. Under the diagonality assumption, the distortion constraint (18) may then be written as

$$\sum_{k=1}^{N} \epsilon (x_k^q(\Theta))^2 w_{kk}(\Theta) 2^{-2R_k(\Theta)} = N. \tag{19}$$

Neglecting the fact that $R_k(\Theta)$ is a nonnegative integer, we can minimize (14) subject to (19) using the Lagrange multiplier method. In particular, its solution is given by

$$\epsilon (x_k^q(\Theta))^2 w_{kk}(\Theta) 2^{-2R_k(\Theta)} = 1 \tag{20}$$

i.e.,

$$R_k^{\text{opt}}(\Theta) = \frac{1}{2} \log_2 \frac{(x_k^q(\Theta))^2 w_{kk}(\Theta)}{C} \tag{21}$$

where the constant $C = 1/\epsilon$. The minimum bit rate then is

$$R_{\min}(\Theta) = \frac{1}{2} \sum_k \log_2 \frac{(x_k^q(\Theta))^2 w_{kk}(\Theta)}{C}. \tag{22}$$

The coefficients for which $(x_k^q(\Theta))^2 w_{kk}(\Theta) < C$ are discarded. For a particular choice of a wavelet, the bit rate requirement may be computed using (22) directly from the transform coefficients. The best wavelet is then identified by minimizing (22) over all vectors Θ.

Let us now make two comments regarding (22). We provide a proof of these comments in the Appendix. First, it may be shown that the coefficients $w_{kk}(\Theta)$ are constant over each wavelet subband. Furthermore, the values of these coefficients depend only on the magnitudes of the frequency responses of the filters used to compute the wavelet decomposition. In Section V, we will argue that in practice it is enough to minimize (22) over all sequences $\{c_k\}$ of length K that correspond to wavelets with $K/2$ vanishing moments. It is clear from the above two comments that if we restrict the optimization problem to these sequences we may take the coefficients $w_{kk}(\Theta)$ in (22) to be constants $w_{kk}(\Theta) = w_{kk}$ independent of Θ. This considerably simplifies the minimization of $R_{\min}(\Theta)$ in (22).

V. A Study of the Optimal Wavelet Coding Wavelet

It is clear from the above discussion that the wavelet based encoding method essentially involves an optimization over all wavelets of a given support length to identify the one that minimizes the bit rate. Before considering this approach further, it is natural to address the following questions:

- Does optimization of the wavelet basis help?
- Do we need to perform a full blown optimization?
- What is the effect of considering longer wavelet sequence $\{c_k\}$ (i.e., wavelets with longer time domain supports)?
- How many levels should we use in a wavelet decomposition?

To answer these questions, we performed several experiments with a wide variety of monophonic music signals sampled at 44.1 kHz. The goal of these experiments was to achieve perceptually transparent coding of several seconds of these signals using the adaptive wavelet encoding technique described in Section IV. We summarize and justify below the conclusions that we have reached based on the results of these experiments.

1) Some optimization is needed: We found that choice of the encoding wavelet significantly affects the achievable level of audio data compression when the goal is to maintain transparent quality. For example, a maximum of 2.5 bits/sample is required to encode the wavelet coefficients (excluding the side information) of a piece of castanets sound when wavelets corresponding to 40 coefficient sequences $\{c_k\}$ are used. On the other hand only 0.8 bits/sample is required when the *optimal* wavelet corresponding to a 40 coefficient sequence $\{c_k\}$ is used in coding the signal.

2) There is no need to perform a full blown optimization in particular when long wavelet sequences $\{c_k\}$ are used: This conclusion is based on the results of the following experiment. For each audio sample we minimized (22) over two different sets of wavelets. In the firsts case we considered all wavelets corresponding to K coefficient sequences $\{c_k\}$. As discussed in Section I, all such wavelets are parametrized by $K/2 - 1$ *angle* parameters. Hence the best wavelet can be identified by performing a grid search over a parameter space of dimension $K/2 - 1$. In the second case, we considered only wavelets that corresponded to K coefficient sequences $\{c_k\}$ and has a maximum number of $K/2$ vanishing moments. We found that

- The typical reduction in bit rate is about 5% when the optimal wavelet with a support of 5 s is used rather than the optimal wavelet with a support of 5 s and 3 vanishing moments. (Wavelets with a support of 5 s correspond to 6 coefficient sequences $\{c_k\}$.)
- The typical reduction in bit rate is about 3.5% when the optimal with a support of 9 s rather than the optimal wavelet with a support of 9 s and 5 vanishing moments. (Wavelets with a support of 9 s correspond to 10 coefficient sequences $\{c_k\}$.)

We have also encoded a more limited number of audio samples using longer sequences $\{c_k\}$. For $K \geq 15$, we observed even smaller reduction in bit rates when the optimal wavelet of support $K - 1$ (corresponding to a K coefficient sequence $\{c_k\}$) is used rather than the optimal wavelet of support $K - 1$ that has $K/2$ vanishing moments. Hence, we conclude that in minimizing (22) over all K coefficient sequences $\{c_k\}$, optimum or near optimum bit rates can be achieved if we restrict the search to sequences that correspond to wavelets with $K/2$ vanishing moments. This is specially true when K is large.

3) Longer sequences $\{c_k\}$ yield better results: This conclusion is not surprising since longer sequences $\{c_k\}$ correspond to wavelet filter banks with sharper transition bandwidths, i.e., to a better separation of frequency information. For a typical audio sample the average bit rate for quantizing the wavelet coefficients was 2.1 bits/sample when an optimal four coefficient wavelet was used. The bit rate went down to 1.75 bits/sample when an optimal 20 coefficient wavelet was used, and to 0.8 bits/sample when an optimal 40 coefficient was used.

4) Deeper Wavelet Decomposition yields better results: Typically, there was about 6% reduction in bit rates when 10 levels were used in the decomposition tree instead of 4. However, it was found that the improvement saturates quickly as the depth of the decomposition tree increases. Note also that perceptual threshold considerations dictate that some wavelet bands be split more than others (c.f. Section I). Because of these considerations, we have chosen to implement a fixed depth decomposition tree in our coder. This fixed decomposition tree is the one shown in Fig. 3.

5) Optimization among wavelets with large number of vanishing moments is helpful: The conclusions that we have listed above indicate that we should use long wavelet sequences $\{c_k\}$ to reduce the bit rate. Furthermore, in optimizing over along sequences $\{c_k\}$ we may limit the search to wavelets possessing a maximum number of vanishing moments. The natural question then is: is it helpful to search for the best wavelet among all wavelets with $K/2$ vanishing moments that correspond to K coefficient sequences $\{c_k\}$? To answer this question we conducted experiments with three sets of wavelets. These were wavelets with 5 vanishing moments that corresponded to 10 coefficient sequences $\{c_k\}$, wavelets with 10 vanishing moments that corresponded to 20 coefficient sequences $\{c_k\}$ and wavelets with 20 vanishing moments that corresponded to 40 coefficient sequences $\{c_k\}$. The cardinality of these three sets is respectively 4, 32, and 1024 (c.f. Section II). For each of these sets, we computed the probability $p(x)$ of obtaining more than x percent difference between the lowest and highest bit rates that can be achieved with wavelets in the set. Fig. 6 illustrates the value of $p(x)$ as a function of x for each of the three sets. It is evident from these plots that both the average gain in bit rates (i.e., the value of x for which $p(x) \simeq 0.5$) and the minimum gain in bit rates (i.e., the value of x for which $p(x) \simeq 1.0$) increase with the size of the support of the wavelet, i.e., with the size of the sequence $\{c_k\}$ corresponding to the wavelet. Furthermore, these values are quite substantial for long wavelets. For example, there is a minimum optimization gain of about 7% for the set of 1024 40 coefficient-20 vanishing moments wavelets. The average gain for that set is on the order of 8.75%. The average gain is even higher when wavelets with longer supports are used (e.g., 15% when 60 coefficient-

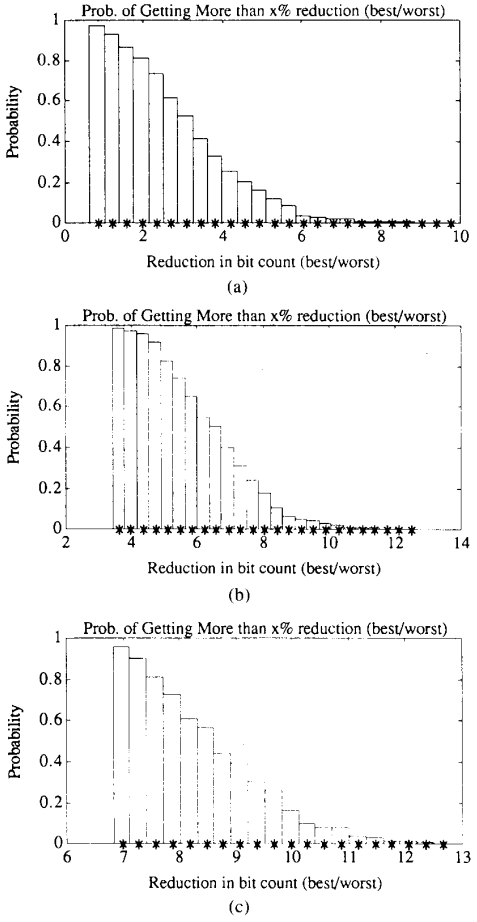


Fig. 6. Probability of getting more than x% reduction in bit count (best/worst-case) when optimization was performed: (a) Among 10 coefficient—5 vanishing moment sequences; (b) 20 coefficient—10 vanishing moment sequences; and, (c) 40 coefficient—20 vanishing moment sequences.

30 vanishing moments wavelets are used). Hence, we conclude that a search for the best wavelet among the set of wavelets of a given support and a maximum number of vanishing moments contributes significantly towards our goal of bit rate reduction.[1]

VI. Dynamic Dictionary Based Encoding

It was noted earlier that bit rate reduction in an audio coding method is possible by exploiting the masking characteristics of human hearing as well as statistical redundancies of the signal. The wavelet based coding method is able to exploit the masking characteristics of human hearing. Further reduction in the bit rate requires getting rid of statistical redundancies in the signal. One possible way to do that is to use predictive coding (differential PCM). However, this method has generally been found to be unsuitable for maintaining transparent quality across a wide range of music signals. Another possibility is to use vector quantization (VQ) [13], [18]. It has been shown that VQ coders are theoretically superior to scalar quantizers in the sense that the rate distortion bounds on the data rate can be approached closer with vector quantization than with scalar quantization. VQ has generally found a limited role in hi-fi audio coding because of lingering doubts about its ability to ensure transparent quality. However, recent work by Chan and Gersho [6] shows that VQ may have a role to play in low bit rate transparent audio coding.

In our work, we have used a simple dynamic dictionary to eliminate statistical redundancies in the signal. The dictionary is maintained by both the encoder and decoder. It is updated at the encoder and decoder using the same set of rules and decoded audio frames. Our approach may be viewed as a crude form of adaptive VQ as illustrated in Fig. 1. Unlike traditional VQ schemes, we transmit the index of the best matching dictionary entry *and* encode the difference between the waveform and that entry using the wavelet based method. This allows us to use a dictionary of relatively small size and rather simple methods for dictionary construction and update.

Specifically, for each frame $\boldsymbol{x}$ of audio data, we first identify the best matching entry $\boldsymbol{x}^D$ currently in the dictionary. Next, we form the residual signal $\boldsymbol{r} = \boldsymbol{x}^D - \boldsymbol{x}$. Both $\boldsymbol{x}$ and $\boldsymbol{r}$ are then encoded using the wavelet based method. Finally, we pick up the code which requires the smaller number of bits for transmission. If we elect to transmit a coded version of $\boldsymbol{r}$, we also transmit the index of the best matching dictionary entry. Note, that the difference signal $\boldsymbol{r}$ may be encoded using the perceptual threshold of the signal itself. The audio signal $\boldsymbol{x}$ is given by $\boldsymbol{x} = \boldsymbol{x}^D + \boldsymbol{r}$. Therefore, the coding error in $\boldsymbol{x}$ is equal to the error in the quantization of the residual $\boldsymbol{r}$. Hence, we simply need to ensure that the quantization error in $\boldsymbol{r}$ has a spectrum that falls below the perceptual threshold estimates corresponding to $\boldsymbol{x}$.

The proposed coding scheme is illustrated in Fig. 7.

A. Dynamic Dictionary Speech

Our dictionary encoding procedure works as follows. Each frame $\boldsymbol{x}$ of audio data is split into two halves z_1 and z_2 (unless the frame length has already been halved for pre-echo control as explained below). For either of the two half frames the dictionary is searched for the best perceptual matches using the procedure described below. The best entries are then subtracted from the respective halves of the signal $\boldsymbol{x}$ to form the residual vector $\boldsymbol{d}$.

To encode a half-frame z_n, $n = 1, 2$, of $N/2$ samples we search through the dictionary to find an element μ_0 which is closest to z_n in terms of a "perceptual distance measure." Even though the dictionary is used to encode vectors of $N/2$ samples its entries are maintained as "meta vectors" of N samples each. This provides us with the ability to optimally align each subframe with each of

[1]In a variant of the proposed coding technique in which one uses different wavelets to compute the transform coefficients in different wavelet subbands, the average gain is on the order of 20–25/when 40 coefficient—20 vanishing moments wavelets are used.

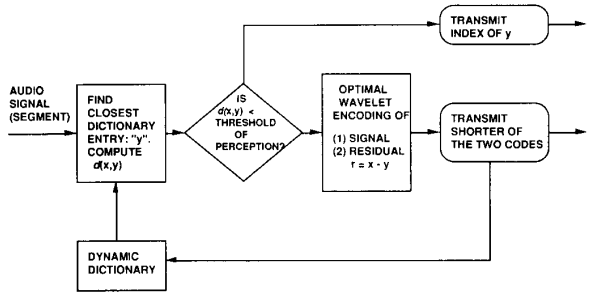


Fig. 7. Dynamic dictionary based encoding of audio signals.

the dictionary entries before computing a perceptual distance measure between the subframe and each dictionary entry. In particular, the perceptual distance measure between the subframe and each dictionary entry is found by executing the following three steps:

1) Compute a sliding window correlation between z_n and μ and find the lag L_i corresponding to the peak of the correlation function ($0 \leq i \leq N/2$). Next, form a vector y_0 consisting of $N/2$ samples of μ starting from lag L_i.

2) Estimate the time warping factor to normalize the time scale of y_0 to that of z_n (see below).

3) Compute a frequency weighted error for the error vector $e = z_n - y_0$ using the perceptual threshold computed from z_n. This error power is the desired perceptual distortion measure.

Steps 1 and 2 above ensure a better perceptual match through an improved time and scale alignment. The incorporation of these steps increases the effective size of the dictionary significantly beyond its physical dimension.

B. Dynamic Dictionary Update

The dictionary in our coding method is *dynamic*, i.e., it is continuously updated based on the stream of incoming audio data. Since the same dictionary is maintained by the encoder and decoder, the dictionary is updated using decoded audio frames. It may be noted that an adaptive dictionary has previously been used, e.g., for the quantization of spectral information in a low bit rate speech coder [27], and for image compression [11]. The dictionary update problem in our coder is somewhat different because dictionary entries consist of waveforms. The vector sizes are therefore relatively large making it impractical to collect several samples before executing an update step. Since our encoder also encodes the residual, we have used the following simple dictionary update procedure.

The minimum distance measure between the decoded signal $\hat{x}$ corresponding to frame x and the perceptually closest entry into the dictionary is compared against a preselected threshold. If it is below the threshold the dictionary remains unchanged. Otherwise the decoded signal $\hat{x}$ is used to update the dictionary using a last-used-first-out (LUFO) strategy, i.e., the last-used entry of the dictionary is replaced by $\hat{x}$. Several improved techniques for dictionary update in the audio coder are currently under investigation.

While the dictionary update procedure as described above is relatively crude when compared to some known techniques used in adaptive VQ [11], it suffices in our case because the residuals are also being transmitted. At the beginning we start with a dictionary that is half-filled with a variety of audio patterns.

VII. Implementation Issues

In this section we discuss some important implementation issues. We begin by discussing the segmentation of audio signals into frames.

A. Adaptive Framing for Preecho Control

In selecting a suitable frame size we have to address two conflicting requirements. A larger frame size is desirable for maintaining lower bit rates. Unfortunately, larger frames sizes also lead to poorer quality because of the nonstationarity of audio signals. The proposed coder employs an analysis/synthesis frame size of 2048 samples (about 46 ms). The two ends of each frame are weighted by the square root of a Hanning window of size 128 (i.e., each two consecutive frames overlap by this amount). While this frame size is usually small enough to maintain good quality for a wide range of music signals, it can lead to significant amount of preechoes in signals containing sudden bursts of energy. The problem of preechoes is countered using an adaptive frame size as described below.

Preechoes arise because of the noncausal nature of the synthesis window. The reconstruction errors are spread over the total effective width of the window in the time domain. These errors will remain inaudible if the signal spectrum is nearly the same over the time duration of the entire window. On the other hand, a sudden increase in the signal energy within a frame leads to a masking threshold that is incorrectly too high for most of the time duration of the window. It is accurate only over that part of the window where the sudden burst of energy has occurred. A study of time domain masking effects indicates that forward masking of an impulse is sufficient to mask the errors that result from using that high incorrect masking threshold in all parts of the window that follow the onset of the burst of energy. Unfortunately, backward masking lasts for a maximum of 4 ms and is not enough to mask errors in the part of the window that precedes the onset of the burst of energy.

Our experiments with sounds specially prone to preecho effects (e.g., castanets) suggest that reducing the frame size to 1024, as opposed to the usual size of 2048, reduces (but does not completely eliminate) the preechoes to a great extent. We therefore need a mechanism for identifying signal frames with sudden bursts of energy.

A sophisticated approach to adaptive frame is discussed in [3]. In our coding scheme framing is done using an

energy-entropy criterion. The entropy is computed by dividing each frame into segments of K ($K = 16$ is a suitable value at the 44.1 kHz sampling rate) samples each. The signal energy is computed over each of these segments and normalized by the overall frame energy. Let us denote by σ_i the normalized energy of the ith short segment. Furthermore, let there be a total of J short segments in a frame with N samples. The entropy of the frame is computed as follows:

$$I = -\sum_{i=1}^{J} \sigma_i^2 \log_2 \sigma_i^2. \tag{23}$$

This entropy measure is similar to probabilistic entropy. In a frame with nearly constant energy the entropy is largest. On the other hand, in frames with sudden energy transitions the value of the entropy measure falls down. The entropy criterion was found to be more robust than a pure energy criterion in our experiments. Fig. 8 shows the plot of a segment of castanets sound and corresponding entropy measure. In our encoder, a *change* in entropy of 1.25 bits or more triggers a switching of the frame size.

We are currently pursuing a better approach to adaptive framing and preecho control. This framing approach also uses the entropy criterion (23). However, rather than simply switching to a shorter frame, it adaptively segments the audio signal to guarantee that any sudden burst of signal energy appear only at the beginning of a frame.

B. Time Warp Factor Estimation

In the dictionary based coding method discussed above we obtain bit rate reductions by transmitting the difference between the signal and a perceptually close dictionary entry. It may be difficult to identify a good dictionary entry (and hence difficult to achieve lower bit rates) because of time scale shifts in the audio signal. What this means is that we cannot take full advantage of situations where the same sound is repeated but with a different time scale. This is a problem that also inhibits speech/speaker recognition systems where it is usually handled by sophisticated dynamic time warping algorithms, e.g., [29]. The approach there is to achieve time-normalization by computing a time-warping function.

In the audio coding application it is expensive to transmit a time-varying warp function. Moreover, an exact characterization of the warping function is not as important in our audio coding method *because residuals are also transmitted*. Therefore, we compute estimates of the warp function over several disjoint segments of each frame of audio data and use the *average* of these estimates for the entire duration of the frame. With this assumption, we can use the following simple approach to estimate this warp factor. Let us denote the warp factor by α. α is chosen to minimize

$$C(\alpha) = \int [x(t) - y(\alpha t)]^2 \, dt \tag{24}$$

where $x(t)$ is the audio signal, $y(t)$ is the dictionary entry under consideration and the integral is performed over a segment of a frame. We assume that these two signals have been normalized to have equal energy.

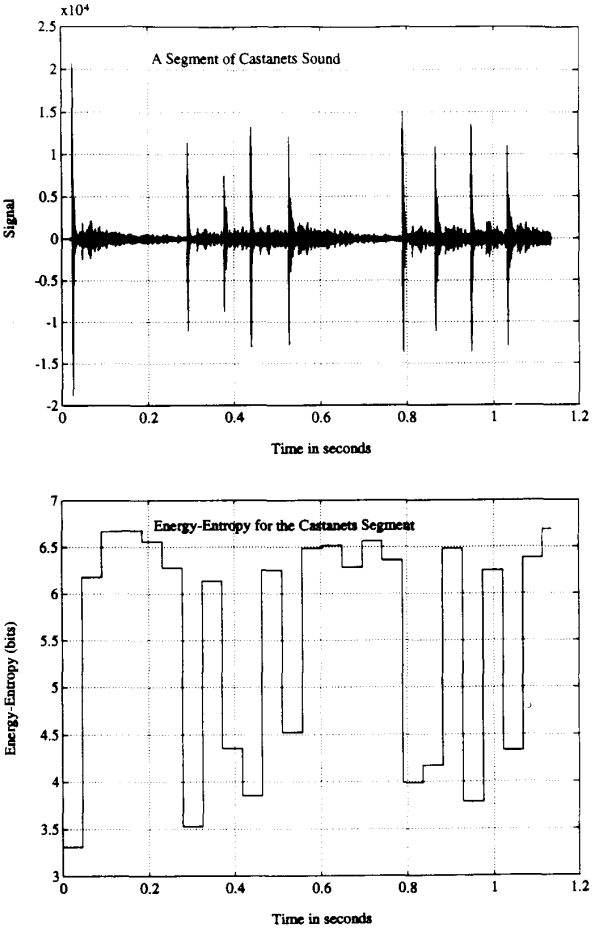


Fig. 8. Plot of the energy-entropy for a short segment of castanets sound.

We restrict the warp factor to be of the form $\alpha = 1 + \gamma$, where $-0.5 \leq \gamma \leq 0.5$. Furthermore, we assume that $y(t)$ is smooth and has a Taylor series expansion at every point. Expanding $y(t)$ and ignoring higher order terms we obtain

$$C(\gamma) = \int [x(t) - y(t) - \gamma t y'(t)]^2 \, dt. \tag{25}$$

We can now find γ by minimizing the cost function in (25). In particular,

$$\gamma = \frac{\int [x(t) - y(t)]\, t y'(t) \, dt}{\int [t y'(t)]^2 \, dt}. \tag{26}$$

For discrete data we can approximate the expression in (26) by

$$\gamma = \frac{\sum_k k[x(k) - y(k)] \cdot [(y(k+1) - y(k)]}{\sum_k k^2 [y(k+1) - y(k)]^2}. \tag{27}$$

In (27) the summations are for all the samples in a segment of a frame.

C. Quantization of Side Information

To reconstruct any given frame, the decoder needs to know the quantized values of all wavelet coefficients corresponding to the frame as well as side information. The side information in the wavelet based coding method consists of the parameters indicating the optimum choice of wavelet, the wavelet domain perceptual threshold $\{w_{kk}\}_{k=1}^{k=N}$, and the dynamic range of each adaptive quantizer. In addition, the signal energy is also transmitted once every frame. Typically, each frame is normalized by its energy before analysis.

Transmission of wavelet parameters is inexpensive because we limit our search to wavelets of a given fixed support and a maximal number of vanishing moments. For example, if we use wavelets with 20 vanishing moments that correspond to 40 coefficient sequences $\{c_k\}$ we need only 10 bits/frame to identify the optimal wavelet. The bit requirement goes up to 15 bits/frame if we use wavelets with 30 vanishing moments that correspond to 60 coefficient sequences $\{c_k\}$.

Transmitting the values of the wavelet domain perceptual thresholds $\{w_{kk}\}_{k=1}^{k=N}$ also turns out to be rather inexpensive. Specifically, Proposition A.1 states that the wavelet domain perceptual thresholds are constant over each wavelet subband. Thus, the number of values of w_{kk} that need to be transmitted per frame is equal to the number of sub-bands in the wavelet decomposition. We quantize these values on a logarithmic scale using 6 bits each.

The quantization of the dynamic ranges of the adaptive quantizers is a somewhat more complex issue. To keep the overhead low, the dynamic ranges are transmitted only once for every group of L wavelet coefficients. While a smaller value for L reduces the bit rate estimates in (22), it can increase the overhead prohibitively. A suitable value for L is best determined experimentally. In our work with audio data sampled at 44.1 kHz, we typically use a value of $L = 8$ and group wavelet coefficients in a time-consecutive fashion. It is advantageous to consider differential coding schemes for the dynamic ranges. This requires a careful investigation of the statistics of the values of the dynamic ranges of the adaptive quantizers. Our studies in this regard are still preliminary. Presently we use the following simple method for quantizing these values. The mean of the logarithms of the dynamic ranges of the adaptive quantizers used in each of the wavelet sub-bands is quantized using 6 bits. We also quantize the difference between the logarithm of each of these dynamic ranges and the mean of their logarithms using a 2 bit nonlinear quantizer optimized for a Gaussian distribution. The variance of the logarithms of the dynamic ranges in each wavelet sub-band is also transmitted. This requires an additional 6 bits. In addition, if preecho control is in effect (i.e., if a shorter 1024 sample frame is used) we quantize the 10 percent of the logarithms of the dynamic ranges that have the largest magnitudes separately using a 6 bit uniform quantizer. This helps in the accurate reconstruction of sharp attacks. All the uniform quantizers of the logarithms of the dynamic ranges employ a step size of 1.1 dB.

The total overhead for the side information is approximately 0.4 bits/sample for audio data sampled at 44.1 kHz. Note, that dynamic dictionary based coding method requires additional side information. This information includes a flag to indicate whether signal or residual is being transmitted, an index for the dictionary entry, a translation parameter, and a time warp factor. Quantization of this additional information adds insigificantly to the overhead since it needs to be transmitted only once per frame.

VIII. Experimental Results

Let us now discuss some subjective and objective testing results that we have obtained in our quest to assess the quality of signals encoded with the proposed coding scheme. In the sequel we shall refer to the proposed coding approach as the wavelet technique coder (WTC). We begin with a discussion of the test material (music samples) that we have used in the tests. Following this, we present the results of the actual tests that we have performed. As we shall see, these results clearly establish the potential of the proposed coder for transparent or near transparent audio compression for a variety of audio material.

A. Test Material

The basic set of audio source material that we used in the analysis and quality assessment of the proposed coding scheme is listed in Table I. All the audio segments in this set are of CD quality. It may also be noted that the set contains some music signals which have been traditionally considered to be "hard" to encode [20], e.g., the castanets, drums, etc. The castanets sound is specially noteworthy because it consists of impulsive energy bursts, or "sharp attacks," followed by long periods of very little signal energy. Such signals not only require larger bit rates but are also extremely susceptible to the "preecho" effect.

B. Subjective Listening Tests

While the opinion of a few "expert" listeners can be quite useful in assessing the quality of a given audio coding approach, one should also seek the opinion of a large group of listeners. The opinion of these listeners can then be quantified in a statistical sense.

It is important to eliminate the element of chance in listening tests that involve transparent coding, where there is little difference between the original and reconstructed signal. Therefore, we should present several stimuli of each source material to each listener. It is also useful not to reveal to the listener the actual order in which the stimuli are presented (e.g., original, coder 1, coder 2, etc.).

TABLE I
A LIST OF SOURCE MATERIAL USED IN THE LISTENING TEST

20 kHz source material sampled at 44.1 kHz, 16 bit PCM	
Code	Instrument/Style
Drum	solo modern drum
Pop	female vocal pop song
Vivc	violin with orchestra
Castanets	solo castanets
Clarinet	solo clarinet
Piano	solo piano

Blind tests are therefore commonly used. One of the most popular blind tests used is the "pair test" or the "double blind" test in which each test sample consists of a pair of stimuli (of the same source materials) and the listener is asked to select the one he or she finds better in overall quality. The test data is then averaged for each stimulus. Ideally, an individual statistical confidence measure for each subject should also be estimated. This is because different test subjects have different hearing experiences and hearing abilities and some problems may be audible only to a few very critical listeners.

The subjective listening tests that we performed with the proposed coding scheme had two goals:

(i) test for transparency (as described above) of the monophonic coding at bit rates of approximately 64 kb/s;

(ii) compare the quality of 64 kb/s coding using the proposed scheme with that of the MPEG layer-2 [3], [25] coding algorithm at 64 kb/s. The MPEG layer-2 coded material was provided to us by Texas Instruments Inc.

In the first set of tests that we performed, we presented audio samples coded at 64 kb/s to an "expert listener." This "expert listener" found the quality of the encoded samples to be "virtually indistinguishable" from the original source material. Furthermore, he judged the quality of encoded signals using the proposed scheme to be clearly better than that of the MPEG layer-2 coding algorithm. He could perceive a degradation in the form of "band limiting" and of distortion in sharp attacks in the MPEG layer-2 coded material. No such degradation was noticed with the proposed coder. However, he did notice some negligible amount of weak preechoes in the coded castanets piece.

The second set of listening tests that we performed involved a group of nine people that were requested to volunteer for subjective evaluation of the proposed coding scheme. Most of the individuals in this group came from the communication and signal processing research groups at the University of Minnesota. The group consisted of eight males and one female listener. All the volunteers in this group had some exposure to signal compression problems. Their ages ranged from 23 to 31 years. They also had a wide range of musical interests including classical, modern, rock and jazz music.

The format of this second set of subjective tests was as follows. The listeners were presented with a total of 38 audio pairs. Each pair was composed of a single source material and had the following format:

- 8–10 seconds of music stimulus (#1)
- 5 seconds of silence
- 8–10 seconds of music stimulus (#2)
- 10–15 seconds of silence.

The listeners were told that either one, both, or none of the stimuli in a pair may be an encoded sample and that more than one coding algorithm had been used. They were then asked to identify the stimulus which they found to be better in overall quality. A "not sure" response was permitted. The responses of the listeners were averaged for each audio source. Table II summarizes the results of the transparency tests. In particular, column 2 shows the probability that the original music sample in column 1 was preferred over its encoded version using the proposed order. Coder quality is transparent (for a piece of music) if the probability in the particular row of Table II is close to 0.5 (i.e., the chance level). The reader may be cautioned that the trial size for these tests is relatively small. Hence, the quantified average probabilities have only limited confidence levels. Nevertheless, these figures clearly demonstrate that the coder provided a transparent or nearly transparent coding for all but one audio source. The quality of the piano signal encoded with the proposed coding approach was not as good as that of other audio pieces encoded with our approach. The piano sample contains long segments of nearly steady or slowly decaying sinusoids. The wavelet based coder does not seem to handle steady sinusoids as well as other signals. It needs to be further optimized for such signals.

Table III on the other hand demonstrates that the wavelet based coder is clearly preferred over the MPEG layer-2 coding of castanets signals at similar bit rates. The results for the piano piece were a virtual tie.

C. Objective Quality Measures

It is well known that the traditionally popular signal-to-noise ratio (SNR) measure has little perceptual meaning and is thus not a good measure of a coder performance. Several attempts to develop objective measures which incorporate psycho acoustic hearing models have been reported. One of the more popular measure is the noise-to-mask ratio (NMR). This measure is reported to be close to the results of the listening tests [20]. However, the main problem with such a measure is that its performance depends on the accuracy of the masking models used in a particular coder. Therefore, it is not easy to compare two different coding algorithms employing different masking models. Other measures have also appeared in the literature and suffer from the same limitation.

In our work, we have elected to use the segmental SNR measure to provide an objective indication of the performance of the decoder. The segmental SNR is measured by computing a SNR measure for each audio frame and averaging these measures over the entire signal. It is more

TABLE II
SUBJECTIVE LISTENING TEST RESULTS: TRANSPARENCY TEST

Music Sample	Average Probability of Original Music Preferred over WTC Encoded Music	Sample Size	Comments
Drums	0.44	18	Transparent
Pop	0.58	36	Transparent
Vivc	0.50	18	Transparent
Castanets	0.61	36	Nearly Transparent
Clarinet	0.53	36	Transparent
Piano	0.66	18	Original preferred

TABLE III
SUBJECTIVE LISTENING TEST RESULTS: COMPARISON WITH MPEG CODING

Music Sample	Average Probability of MPEG Layer-2 Encoded Music Preferred over WTC Encoded Music	Sample Size	Comments
Castanets	0.33	45	WTC clearly preferred
Piano	0.53	36	Same quality

TABLE IV
SEGMENTAL SNR VALUES FOR THE TEST SAMPLES

Sample	SEG-SNR (dB)
Castanets	26
Piano	23
Pop	25
Drum	28
Clarinet	25
Vivc	21

reliable than a single global SNR computed for a whole audio signal. Table IV lists the segmental SNR values associated to the various audio pieces that we have encoded.

IX. CONCLUSION

We have presented a new audio coding method based on adaptive optimal wavelet basis selection and dynamic dictionary encoding procedures. We also developed methods for incorporating results from the psycho acoustic studies of perceptual masking into the adaptive optimal wavelet coder. Our studies indicate that optimization of the wavelet basis to match the audio data clearly results in a significant reduction in the bit rate requirement. In general, it is advantageous to use longer wavelets, as these provide higher compression. However, it is not necessary to perform a full blown optimization to identify the best wavelet. Optimization among wavelets with large number of vanishing moments yields near optimal results.

Several improvements in the proposed method are possible both in terms of reducing its computational complexity and its bit rate requirements. There are two main computational burdens in the proposed coding method: identification of the optimal wavelet and dictionary search. In the former case, a fast optimization method should be feasible. This is because the optimization may be restricted to all wavelets with a large number of vanishing moments. All wavelets of a given support and with an equal number of vanishing moments correspond to filter banks that have identical magnitude responses but different delays. It follows from Section IV that the only factor affecting the bit rates in this case is the set of peak values of the transform coefficients. The bit rates that can be achieved with all such wavelets would have been identical if no decimation was involved in the wavelet filter banks. On the other hand, these peak values are not shift invariant under the operation of decimation. The optimization therefore seems to be one of finding the filter delays that best exploit the (fixed) decimation scheme at hand. We are currently working on identifying the best wavelet based on this idea. We are also investigating faster search techniques for the dictionary encoding which is the other main source of computational complexity in our scheme.

The bit rates achievable with the proposed method can be further reduced using the following approaches. First, we note that the model for perceptual masking used in this paper is a relatively simple one. Some of the recently published studies introduce more sophisticated masking models and criteria [9], [38]. These studies could lead to further reduction in bit rates. Secondly, the side information currently amounts for about 40% of the bit rate requirements. Much can still be done for the quantization of the dynamic ranges of the adaptive quantizers. Suitable transformations should be considered for the quantization of other parameters. Many of these issues involve further studies of the statistics of the signals and the parameters. This will be one of the focus of our future studies. The most promising approach to bit reduction undoubtedly involves the joint (vector) quantization of groups of wavelet coefficients. Finally, we note that encoded wavelet transform coefficients may still contain redundancies which can be exploited using an entropy coding method, e.g., a Huffman code [18] or a Ziv–Lempel type of encoding.

APPENDIX

Proposition A.1: Let $\boldsymbol{Q}$ be an $(N \times N)$ wavelet transform matrix and $\boldsymbol{W}$ be the $(N \times N)$ discrete Fourier transform matrix. Furthermore, let $\boldsymbol{R_D}$ be the diagonal matrix described in Section III. The entries of $\boldsymbol{R_D}$ are the discretized values of $1/S_n(f)$ and are constant over the frequency domain support of each wavelet sub-band. Then,

1. the diagonal entries of the $N \times N$ matrix $\boldsymbol{QW'R_DWQ'}$ are constant over each wavelet subband,
2. the values of the diagonal entries of the $N \times N$ matrix $\boldsymbol{QW'R_DWQ'}$ depend only on the magnitude of the discrete time Fourier transform of the sequence $\{c_k\}$ that corresponds to the matrix $\boldsymbol{Q}$,
3. the matrix $\boldsymbol{QW'R_DWQ'}$ is nearly diagonal whenever the matrix $\boldsymbol{Q}$ corresponds to a wavelet with a large number of vanishing moments p.

Proof: 1 and 2.

Note that (k, l)th element of $\boldsymbol{QW'R_DWQ'}$ is given by

$$\rho_{kl} = \sum_{i=0}^{N-1} r_{ii} P_k(\omega_i) P_l^*(\omega_i) \tag{28}$$

where $P_k(\omega_i)$, denotes the (k, i)th element of the matrix $\boldsymbol{QW'}$ and r_{ii} denotes the (i, i)th of the diagonal matrix $\boldsymbol{R_D}$. Note, that $P_k(\omega_i)$ is the Fourier transform of the kth row of $\boldsymbol{Q}$ evaluated at a frequency $\omega_i = (2\pi i/N)$. When $k = l$ we have,

$$\rho_{kk} = \sum_{i=0}^{N-1} r_{ii} |P_k(\omega_i)|^2. \tag{29}$$

For a particular wavelet subband, the corresponding rows of $\boldsymbol{Q}$ form a circulant block. Hence, the magnitude of the Fourier transforms of any two rows in $\boldsymbol{Q}$ which correspond to the same sub-band are equal at each frequencies; i.e., $|P_k(\omega_i)| = |P_l(\omega_i)|$ $\forall i$ when k and l belong to the same wavelet subband. Therefore, from (29), $\rho_{kk} = \rho_{ll}$, when indexes k and l correspond to wavelet transform coefficients in the same subband. Finally, note that ρ_{kk} depends only on $|P_k(\omega_i)|$ which establishes 2.

3. We will assume in what follows that wavelet coefficients have been normalized properly so that $\Sigma_{i=0}^{N-1} |P_k(\omega_i)|^2 = 1$ $\forall k$.

First, note that by assumption the matrix $\boldsymbol{R_D}$ is diagonal and has entries that are constant over the frequency domain support of each wavelet subband. Thus we may write $\boldsymbol{R_D}$ as a sum of diagonal matrices $\boldsymbol{D_j}$. Each matrix $\boldsymbol{D_j}$ is zero except for its diagonal entries that correspond to the frequency domain support of the jth wavelet subband. Furthermore, all the nonzero diagonal entries of $\boldsymbol{D_j}$ are equal to a single value d_j.

Consider now the matrix $\boldsymbol{QW'D_jWQ'}$. The (k, l)th element of this matrix is given by

$$r(k, l) = d_j \sum_{\omega_i \in S_j} P_k(\omega_i) P_l^*(\omega_i) \tag{30}$$

where S_j is the set of frequencies ω_i that belong to the frequency domain support of the jth wavelet subband. When $k = l$ we obtain

$$r(k, k) = d_j \sum_{\omega_i \in S_j} |P_k(\omega_i)|^2. \tag{31}$$

If k belongs to the jth wavelet subband we may use the fact that $\Sigma_{i=0}^{N-1} |P_k(\omega_i)|^2 = 1$ $\forall k$ to write (31) as

$$r(k, k) = d_j - d_j \sum_{\omega_i \notin S_j} |P_k(\omega_i)|^2. \tag{32}$$

The second summation in the equation above is essentially the stop band power of the wavelet filter $P_k(\omega)$. It is clear from Fig. 2(a) (plots of transfer function magnitude of wavelet filters with large number of vanishing moments) that the stop-band power decreases to zero as $p \to \infty$.

Let us now derive an upper bound on that stop-band power. It may be shown that the stop-band power of any wavelet filter is smaller than twice that of the first low pass filter in the decomposition. Assume now that the wavelet used in the decomposition corresponds to a sequence $\{c_k\}$ of length $2p$ and that it possesses exactly p vanishing moments. Using [8, Proposition 4.5, p. 977] and the definition of $P(\omega)$ given in Section II we find that the stop-band power of the first low pass filter in the decomposition corresponding to such a wavelet is

$$P^{\text{stop-band}} \le 2 \int_{\pi/2}^{\pi} |P(\omega)|^2 \, d\omega \tag{33}$$

$$\le 2 \int_{\pi/2}^{\pi} \cos^{2P}\left(\frac{\omega}{2}\right) \sum_{k=0}^{P-1} \binom{P-1+k}{k} \cdot \sin^{2k}\left(\frac{\omega}{2}\right) d\omega$$

$$\le 2 \sum_{k=0}^{P-1} \binom{P-1+k}{k} \cdot \int_{\pi/4}^{\pi/2} \cos^{2P}(\omega) \sin^{2k}(\omega) \, d\omega. \tag{34}$$

Using the well known expressions for the integrals $\int \cos^{2P} \omega \sin^{2k} \omega \, d\omega$ and $\int \cos^{2P} \omega \, d\omega$ we obtain

$$P^{\text{stop-band}} \le Cm(p)v(p) \equiv \epsilon(p) \tag{35}$$

where C is a constant nearly equal to 2 and

$$m(p) = \left[\frac{\pi}{4} - \frac{1}{2} \sum_{k=1}^{P-1} \frac{(k-1)!}{1.3.5 \cdots 2k-1)}\right] \tag{36}$$

$$v(p) = \prod_{i=1}^{p} \left(1 - \frac{1}{2i}\right). \tag{37}$$

Both $m(l)$ and $v(l)$ are small constants which decay with increasing p. The decay of $m(p)$ is faster than that of $v(p)$. Hence, we have from (32)

$$|r(k, k)| - d_j| \le d_j \epsilon(p). \tag{38}$$

For $p = 20$, $\epsilon(p) \sim 10^{-6}$.

On the other hand, if k does not belong to the jth wavelet subband we may use an analysis similar to the one that we presented above to deduce that

$$|r(k, k)| \le d_j \epsilon(p). \tag{39}$$

Next, we consider the case where $k \neq l$ but k and l belong to the same wavelet subband. Note that the block circulant structure of $\boldsymbol{Q}$ indicates that in this case

$$|P_k(\omega_i)| = |P_l(\omega_i)| \qquad \forall i = 0, \cdots, N-1. \tag{40}$$

Furthermore, (30) implies that

$$r(k, l) = d_j \sum_{i=0}^{N-1} P_k(\omega_i) P_l^*(\omega_i) - d_j \sum_{\omega_i \notin S_j} P_k(\omega_i) P_l^*(\omega_i). \tag{41}$$

Since $\boldsymbol{QW'WQ} = \boldsymbol{I}$ where $\boldsymbol{I}$ is the $N \times N$ identity matrix

we have

$$\sum_{i=0}^{N-1} P_k(\omega_i) P_l^*(\omega_i) = 0. \tag{42}$$

Substituting this into (41) and using (40) above we obtain

$$|r(k, l)| \leq d_j \sum_{\omega_i \notin S_j} |P_k(\omega_i)|^2. \tag{43}$$

The sum on the right is once again and d_j times the stopband power. Hence,

$$|r(k, l)| \leq d_j \epsilon(p). \tag{44}$$

Finally, we consider the case where k and l belong to different wavelet bands. In this case (42) remains valid, but equality (40) is no more true. However, we note that

$$|r(k, l) \leq d_j \sum_{\omega_i \notin S_j} |P_k(\omega_i) P_l^*(\omega_i)| \tag{45}$$

$$\leq d_j \left(\sum_{\omega_i \notin S_j} |P_k(\omega_i)|^2 \right)^{1/2} \left(\sum_{\omega_i \notin S_j} |P_l(\omega_i)|^2 \right)^{1/2}. \tag{46}$$

If either k or l belongs to the jth wavelet subband we still have

$$|r(k, l)| \leq d_j \sqrt{\epsilon(p)}. \tag{47}$$

Otherwise, we have

$$|r(k, l) \leq d_j \epsilon(p). \tag{48}$$

Hence, we conclude that the (k, l) entry ρ_{kl} of $\boldsymbol{QW' R_D WQ'}$ satisfies

$$d_j(1 - \epsilon(p)) \leq \rho_{kk} \qquad (k, k) \text{ belongs to the } j\text{th wavelet subband}$$

$$|\rho_{kl}| \leq \left(\sum_i d_i \right) \epsilon(p) \qquad k \text{ and } l \text{ belong to the same wavelet subband}$$

$$|\rho_{kl}| \leq (d_j + d_i) \sqrt{\epsilon(p)} + \left(\sum_{m \neq i,j} d_i \right) \epsilon(p)$$

k and l belong respectively to the ith and jth wavelet subbands.

Thus, the only elements ρ_{kl} of $\boldsymbol{QW' R_D WQ'}$ that have nonnegligible magnitude will be those diagonal entries ρ_{kk} for which (k, k) belongs to a wavelet subband that corresponds to a nonnegligible element d_j of $\boldsymbol{R_D}$. □

REFERENCES

[1] B. S. Atal, "Predictive coding of speech at low bit rates," *IEEE Trans. Commun.*, vol. COM-30, pp. 600–614, Apr. 1982.

[2] G. Beylkin, R. Coifman, and V. Rokhlin, "Fast wavelet transforms and numerical algorithms I," *Commun. Pure Appl. Math.*, vol. 44, pp. 141–183, Mar. 1991.

[3] K. Brandenburg, G. Stoll, *et. al.*, "The ISO-MPEG-Audio codec: A generic-standard for coding of high quality digital audio," *J. Audio Eng. Soc.*, preprint.

[4] K. Brandenburg, "Ein beitrag zu den verfahren und der qualitaetsbeurteilung fuer hochwertige musikcodierung," Ph.D. dissertation, Univ. of Erlangen, Erlangen, Germany, 1989.

[5] A. Buzo, A. H. Gray, R. M. Gray, and J. D. Market, "Speech coding based upon vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 562–574, Oct. 1980.

[6] W. Y. Chan and A. Gersho, "High fidelity audio transform coding with vector quantization," in *Proc. ICASSP*, pp. 1109–1112, Ar. 1990.

[7] R. R. Coifman, Y. Meyer, S. Quake, and M. V. Wickerhauser, "Signal processing and compression with wave packets," Dept. Math., Yale Univ., 1990, preprint.

[8] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Commun. Pure Appl. Math.*, vol. 41, pp. 909–996, Nov. 1988.

[9] B. Delgutte, "Physiological mechanisms of psycho physical masking: Observations from auditory nerve fibers," *J. Acoust. Soc. Am.*, pp. 791–813, 1990.

[10] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Boston, MA: Kluwer, 1992.

[11] A. Gersho and M. Yano, "Adaptive vector quantization by progressive codevector replacement," in *Proc. ICASSP*, pp. 133–136.

[12] R. A. Gopinath and C. S. Burrus, "Wavelet transforms and filter banks," in *Wavelets and Applications*, C. H. Chi, Ed. San Diego, CA: Academic, 1992.

[13] R. Gray, "Vector quantization," *IEEE ASSP Mag.*, pp. 4–29, Apr. 1984.

[14] A. Grossman and J. Morlet, "Decomposition of hardy functions into square integrable wavelets of constant shape," *SIAM J. Math.*, vol. 15, pp. 723–736, 1984.

[15] R. P. Hellman, "Asymmetry of masking between tone and noise," *Perception Psychophys.*, vol. 11, no. 3, pp. 241–246, 1981.

[16] P. N. Heller, Regular M-band wavelets, Aware Inc., Cambridge, MA, Tech. Rep., AD920608, 1992.

[17] L. E. Humes and W. Jesteadt, "Models for the additivity of masking," *J. Acoust. Soc. Amer.*, vol. 85, no. 3, pp. 1285–1294, 1989.

[18] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[19] L. A. Jeffires, "Masking," in *Foundations of Modern Auditory Theory*. New York: Academic Press, 1970, pp. 87–114.

[20] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select Areas Commun.*, vol. 6, pp. 314–323, 1988.

[21] ——, "Perceptual transform coding of wideband stereo signals," in *Proc. ICASSP '89*, pp. 1993–1996.

[22] M. A. Krasner, "The critical band coder-digital encoding of speech signals based on the perceptual requirements of the auditory system," in *Proc. ICASSP '80*, Denver, CO, pp. 327–331.

[23] S. Mallat, "Multifrequency channel decomposition of images and wavelet models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 2091–2110, 1989.

[24] Y. Meyer, *Ondelettes et Operateurs*. Paris: Herman, 1990.

[25] H. G. Musmann, "The ISO coding standard," presented at *Proc. GLOBECOM '90*, 1990.

[26] National Communications System, "Details to assist in implementation of federal standard 1016 CELP," Technical Information Bulletin 92-1, Jan. 1992.

[27] D. B. Paul, "A 500–800 bps adaptive VQ vocoder using a perceptually motivated distance measure," in *Proc. IEEE GLOBECOM 82*, pp. 1079–1082, 1982.

[28] D. Pollen, "$SU_I(2, F[z, /z])$ for F a Subfield of C," *J. Amer. Math. Soc.*, vol. 3, pp. 611–624, July 1990.

[29] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 43–48, Feb. 1978.

[30] B. Scharf, "Critical bands," in *Foundations in Modern Auditory Theory*. New York: Acadmic, 1970, pp. 150–202.

[31] S. Singhal, "High quality audio coding using multipulse LPC," in *Proc. ICASSP*, pp. 1101–1104, Apr. 1990.

[32] S. M. F. Smyth and J. V. McCanny, "4 bit hi-fi: Hi quality music coding for the ISDN broadcasting applications," in *Proc. ICASSP '88*, pp. 2532–2535.

[33] G. Stoll and Y. F. Dehery, "High quality audio bit rate reduction system family for different applications," presented at *Proc. Supercomm. '90*, Atlanta, GA, 1990.

[34] G. Strang, "Wavelets and dilation equations: A brief introduction," *SIAM Rev.*, vol. 31, no. 4, pp. 614–627, 1989.

[35] A. H. Tewfik and M. Kim, "Correlation stucture of the wavelet coefficients fo fractional brownian motions," *IEEE Trans. Informat. Theory*, vol. 38, pp. 904–909, Mar. 1992.

[36] ——, "Fast multiscale statistical signal processing algorithms," to appear in, *IEEE Trans. Signal Processing*.

[37] A. H. Tewfik, D. Sinha, and P. Jorgensen, "On the optimal choice of a wavelet for signal representation," *IEEE Trans. Informat. Theory*, vol. 38, pp. 747-765, Mar. 1992.
[38] R. N. J. Veldhius, "Bit rates in audio source coding," *IEEE J. Select Areas Commun.*, vol. 10, pp. 86-96, Jan. 1992.
[39] R. N. J. Veldhuis, M. Breeuwer, and R. G. Van Der Waal, "Subband coding of digital audio signals," *Philips Res. Rep.*, vol. 44, no. 2-3, pp. 329-343, 1989.
[40] P. Voros, "High quality sound coding with 2 × 64 kb/s using spontaneous dynamic bit allocation," in *Proc. ICASSP '88*, pp. 2536-2539.
[41] G. W. Wornell, "A Karhunen-Loeve like expansion for $1/f$ processes via wavelets," *IEEE Trans. Informat. Theory*, vol. 36, pp. 859-861, 1990.
[42] H. Zou and A. H. Tewfik, "Parametrization and compactly supported orthonormal wavelets," *IEEE Trans. Signal Processing*, vol. 41, pp. 1428-1430, Mar. 1993.
[43] ——, "A theory of M-band compactly supported orthonormal wavelets," submitted to, *IEEE Trans. Circuits Syst.*, Feb. 1992.
[44] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *J. Acoust. Soc. Am.*, vol. 33, p. 248, 1961.

Deepen Sinha (S'92) was born in Gonda, India. He received a B.Tech. (Hons.) degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, in 1986, a M.S. degree in electrical engineering from the Iowa State University, Ames, IA in 1989, and a Ph.D. degree in electrical engineering from the University of Minnesota, Minneapolis, MN in 1993.

From 1986-1987, he worked as a computer research engineer at UPTRON in Lucknow, India. He is currently with the Signal Processing Research Department at the AT&T Bell Labs in Murray Hil, NJ. The focus of this work at present is in the area of multichannel audio compression. His other research interests are in the areas of very low bit rate coding of speech and music signals, video compression, time-frequency signal representation, and filterbanks.

Dr. Sinha is a member Phi Kappa Phi.

Ahmed H. Tewfik (S'82-M'87-SM'92), for a photograph and biography, see page 2849 of the September 1993 issue of this TRANSACTIONS.