High-Quality Audio Compression Using an Adaptive Wavelet Packet Decomposition and Psychoacoustic Modeling

Pramila Srinivasan, Member, IEEE, and Leah H. Jamieson, Fellow, IEEE

Abstract—This paper presents a technique to incorporate psychoacoustic models into an adaptive wavelet packet scheme to achieve perceptually transparent compression of high-quality (44.1 kHz) audio signals at about 45 kb/s. The filter bank structure adapts according to psychoacoustic criteria and according to the computational complexity that is available at the decoder. This permits software implementations that can perform according to the computational power available in order to achieve real time coding/decoding. The bit allocation scheme is an adapted zero-tree algorithm that also takes input from the psychoacoustic model. The measure of performance is a quantity called subband perceptual rate, which the filter bank structure adapts to approach the perceptual entropy (PE) as closely as possible. In addition, this method is also amenable to progressive transmission, that is, it can achieve the best quality of reconstruction possible considering the size of the bit stream available at the encoder. The result is a variable-rate compression scheme for high-quality audio that takes into account the allowed computational complexity, the available bit-budget, and the psychoacoustic criteria for transparent coding. This paper thus provides a novel scheme to marry the results in wavelet packets and perceptual coding to construct an algorithm that is well suited to high-quality audio transfer for internet and storage applications.

Index Terms—Audio compression, wavelet applications, wavelet packets, zero trees.

I. INTRODUCTION

The increasing volume of multimedia transfer over the internet has necessitated schemes for efficient signal processing to enable real-time audio and video data compression and decoding. While the applications related to telephony have concentrated on the coding of voice signals, applications of high-quality music storage and transfer have motivated the study of the efficient compression of music. Speech signal compression has traditionally exploited the speech production model to remove redundancies, while compression of high-quality audio tries to exploit the noise masking properties of the human ear to advantage. In this work, we try to combine effective techniques of exploiting the masking property, efficient bit allocation, and progressive transmission. In the context of audio compression, progressive transmission refers to the property that as the size of the bit stream increases, the quality of the decoded audio signal improves. This has particular application in internet transfers. The other key aspect of the proposed scheme is the adaptation to available computational power. Recognizing that various platforms that run the decoding software might have different levels of computational power, the algorithm attempts to adapt the filter bank structure so that the reconstruction algorithm can potentially decode in real time. The proposed method therefore brings together several powerful compression techniques to build a coding scheme for transparent audio at about 45 kb/s. Preliminary perceptual results show that this scheme is promising and widely applicable.

We first review some approaches to different aspects of the problem considered in this paper, thus citing the background to this research. In [1], Fastl determines the masking patterns of critical and broadband noise maskers. In [2], Johnston describes a perceptual transform coder that calculates the noise masking thresholds and uses the absolute threshold information to compress CD or DAT music signals. This work also outlines the concept of perceptual entropy (PE) and plots the perceptual entropy histograms for various sources of audio. In [3], Johnston et al. give an overview of techniques that are most likely to be used in audio compression for various services in the future. Considerable emphasis is on the perceptual entropy. In information theory, we learn that the source can be decoded with zero error if the encoder transmitted at a rate greater than or equal to the entropy. The fundamental limit to which we can compress a signal with zero perceived distortion is a semi-mathematical quantity: the perceptual entropy.

Filter banks [4] have been studied extensively lately in the context of data compression and denoising [5]–[11]. In [12], Sinha and Tewfik detail an optimal wavelet structure selection and dynamic dictionary coding scheme to achieve almost transparent coding of CD-quality signals at bit rates of 48–66 kb/s. This work offers unique insights into the use of adaptive wavelet transforms to audio compression.

Simultaneously, in the area of image compression, rapid strides have been made in compression techniques that involve transform coding and sophisticated bit allocation methods. The abundance of image data transfers over the internet has
prompted techniques to facilitate efficient progressive transmission. One notable work is that of Shapiro [13], which is called the embedded zero-tree wavelet algorithm, which aims at exploiting the structure of wavelet coefficients and transfers the image progressively from coarse to fine resolutions. We have incorporated a one-dimensional (1-D) adaptation of this algorithm in this work, along with suitable modifications to use the psychoacoustic model.

In the sphere of industry standards, the Joint Technical Committee (JTC1) established by the International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), which is also known as the Moving Pictures Experts Group (MPEG), prepared the international standard for the coded representation of moving pictures and associated audio. Part 3 of this document, which describes the coding of audio, will be referred to as MPEG I Audio [14] in this paper. MPEG I Audio is an elaborate description of subband coding, psychoacoustic modeling, and related implementation issues. Pan [15] and Brandenburg [16] describe the intricacies of this specification detailing the different layers of compression and the two psychoacoustic models employed. The three layers refer to varying degrees of complexity and bit rates. Layer I refers to the simplest coder/decoder offering bit rates above 128 kb per channel. Layer II has an intermediate complexity and offers about 128 kb per channel. Layer III, which is the most complex of all, offers around 64 kb per channel.

All three layers of the ISO-MPEG I audio standard use a polyphase filter bank for signal decomposition into 32 equal width subbands. This is a computationally simple solution and provides reasonable time–frequency resolution. However, as noted in [15], this approach has three notable deficiencies. First, the equal subbands do not reflect the critical bands of noise masking. Hence, since a single subband might cover more than one critical band, the quantization error introduced in that subband cannot be tuned exactly to the noise threshold in both critical bands. (Typically, the smaller threshold that yields the more conservative bit allocation is chosen). Second, the filter bank and its inverse do not yield perfect reconstruction. This introduces errors even in the absence of quantization error. Third, adjacent filter banks have overlap; therefore, a single tone can affect two filter banks.

In this paper, we present a novel scheme that marries results in wavelet packets and perceptual coding. Fig. 1 shows a block diagram of the encoder/decoder. In the following sections, we detail the components of the system.

The key issues considered are as follows.

- the subband structure, which determines the computational complexity of the algorithm for each frame;
- the PE of the signal segment;
- the efficiency of the bit allocation scheme, which depends on the temporal resolution of the decomposition

We address these issues by introducing the concept of subband perceptual rate (SUPER), which is a measure that tries to adapt the subband structure to approach the PE as closely as possible.

In the remainder of the paper, we detail the components of the system and present results.
problem formulation in Section III. Section IV presents a description of the wavelet packets approach (from the literature) and our adaptation to solve the problem under consideration. The description of bit allocation using our adaptation of the traditional zero-tree method follows. Section VI summarizes our scheme. In Section VII, we describe the experimental setup and conclude with results and remarks on the scope of the algorithm.

II. THE PSYCHOACOUSTIC MODEL

The psychoacoustic model plays a central role in this compression technique. The model, which is based on the properties of hearing, determines the noise energy that is masked in the neighborhood of a tone (which is called frequency masking). This fact is exploited by the quantizer to allow as much noise (thus saving bits) as can be masked. The psychoacoustic model used in this paper closely resembles Model II of the ISO-MPEG [14] specification. The MPEG I Psychoacoustic Model I identifies tonal components based on local peaks of the audio spectrum. The MPEG I Psychoacoustic Model II uses data from the previous two windows to predict, via linear extrapolation, the component values for the current window. Tonal components, being more predictable, thus have higher tonality indices. We give a qualitative description based on [16].

The psychoacoustic model starts with the frequency domain representation, from which the noise-masking thresholds for the critical bands are calculated. The frequency-domain representation is calculated via a 1024-point FFT. The magnitude-phase representation of the FFT is used to calculate the tonality of the current frame. The tonalit measure, which ranges from 0 to 1, is based on the predictability of the current frame from the past two frames. A frame that differs less from the prediction has a higher tonality index. This index is used later to determine the just-masked noise level.

The magnitude values of the frequency domain representation are converted to a critical band representation and convolved with the spreading function. The spreading function describes the noise-masking property, which is the property of the ear-to-mask noise at a frequency in the neighborhood of a tone. The “just masked” noise level is calculated from this function and the tonality index. It is a convex combination of the “noise masking tone” and the “tone masking noise” thresholds, the factor (between 0 and 1) being the tonality index. The threshold in quiet is incorporated, which significantly lowers bit rate estimates for the higher frequencies. This block calculates the masking threshold and the energy in each subband, given a decomposition structure.

In a perceptual subband coding scheme, in order to achieve transparent coding, the threshold used to determine the quantization noise of a subband (which determines the number of bits required) should be the minimum of the thresholds of the frequency lines (or threshold calculation partitions) that the subband contains. This means that we might encounter a very conservative estimate for the threshold of a wide subband. Hence, the psychoacoustic model is most efficiently used when the subbands are as close as possible to the critical bands. This leads us away from the regular structure of conventional wavelet structures, and hence, we look into the wavelet packet approach [17]–[19].

In this work, we extend the above psychoacoustic model in order to fit into the framework of our adaptive wavelet structure. The main output of our psychoacoustic model block is a measure called the SUPER for a subband structure. The SUPER is used to decide on the need to further decompose the subband. This helps to prevent high estimates of SUPER due to several critical bands with different bit rate requirements coalescing into one subband.

The psychoacoustic model calculates the thresholds for each threshold calculation partition, which are denoted by \( t_i, i = 1 \cdots I \), where \( I \) is the number of such partitions. For transparency, assume that for each subband, the threshold to be considered is the minimum of that of all constituent threshold calculation partitions. However, the boundaries of \( t_j \) may not coincide with the boundaries of the subband partitions denoted by \( s_k, k = 1 \cdots K \), where \( K \) is the number of subband partitions. In that event, the noise-masking threshold is calculated for the subband \( s_k \) as follows. Let

\[
\bigcup_{j=j'}^{j'} t_j \geq s_k.
\]

Then, the threshold of \( s_k \), \( t_{s_k} \), is given by \( \min_j t_j \). Let the energy in subband \( k \) be \( c_k \). This threshold is used to find the mask-to-signal ratio, which is denoted by \( t_{s_k} / c_k \). The threshold is also used to find the SUPER. The SUPER is the iteratively determined minimum number of bits \( \sum_k b_k \), where \( b_k \) is the number of bits per sample in the subband \( k \) so that

\[
\epsilon_{q_k} < t_{s_k}
\]

where \( \epsilon_{q_k} \) is the quantization error incurred due to quantizing using \( b_k \) bits. Although it is computationally efficient to calculate this quantity using statistical properties of the signal in each subband, we have deterministically calculated the quantization error for the given signal in this implementation.

Because of this quantity, perceptually transparent compression is possible, given the subband structure, if the rate is greater than or equal to this quantity that depends on the signal and filter bank structure. Fig. 2 illustrates this concept. Fig. 2(a) shows the thresholds assumed for the example. In Fig. 2(b), it can be observed that the bit rate has to be chosen so that the quantization error is less than the minimum threshold \( \epsilon_1 \). In Fig. 2(c), the subband is divided by further filtering and decimation. The output of subband 0 is encoded, based on the threshold \( \epsilon_1 \), and the output of subband 1 is encoded based on threshold \( \epsilon_4 \). Since \( \epsilon_4 > \epsilon_1 \), the output of subband 1 can tolerate more error, hence, it requires fewer bits per sample for transparent coding.

III. THE PROBLEM FORMULATION

Having adopted a wavelet packet filter bank structure and developed a compatible measure for the psychoacoustic model, we proceed to formalize the problem. The problem that we address in this paper can be stated as follows. Assume
that the number of computations per frame of data is \( c \) (depending on the length of the filter) for one stage of filter-bank decomposition. Assume that the total computations allowed (the minimum of the computations permitted at the encoder and decoder) is \( C \).

Therefore, the problem is to adaptively decide the optimal subband structure \( s_k \) that achieves the minimum \( \text{SUPER} \) (which is limited by the \( \text{PE} \)), given the maximum computational limit \( C \) and the best temporal resolution possible (that renders the bit allocation scheme most efficient).

The following section describes the wavelet packet structure and how it can be used to solve the above problem algorithmically.

IV. WAVELET PackETS AND INCORPORATING COMPUTATIONAL COST AND PERCEPTUAL CRITERIA

In [17], [19], [20] and other works, the authors have considered an adaptive filter bank structure to implement signal compression and synthesis schemes. This is an implementation wherein a choice of basis can be made based on relevant criteria. One possible choice is the traditional wavelet basis. The following discussion follows the notation of [20] and [21] and extends this formulation to incorporate perceptual criteria in order to solve the problem under consideration.

In order to define notation, we briefly review the theory of the construction of a wavelet packet library. We start with a quadrature mirror filter \( h(n) \) satisfying

\[
\sum_n h(n-2k)h(n-2l) = \delta_{n,l},
\]

\[
\sum_n h(n) = \sqrt{2}.
\]

Letting \( g_k = (-1)^k h_{1-k} \), we define the following operators \( F_0 \) on \( \mathcal{F}(\mathbb{Z}) \) into \( \mathcal{F}(\mathbb{Z}) \)

\[
F_0 s_k(2i) = \sum_k s_k h_{k-2i},
\]

\[
F_1 s_k(2i) = \sum_k s_k h_{k-2i}.
\]

The map \( F: \mathcal{F}(\mathbb{Z}) \rightarrow \mathcal{F}(\mathbb{Z}) \oplus \mathcal{F}(\mathbb{Z}) \) defined by \( \mathbf{Z} = F_0 \oplus F_1 \) is orthogonal. In addition

\[
F_0 F_0^* = F_1 F_1^* = I,
\]

\[
F_0^* F_0 + F_1^* F_1 = I.
\]

Starting from a function \( W_0(x) \) that is a scaling function \( \phi \) and \( W_2(x) \), which is a wavelet function \( \psi \), we define the following series of functions (as in [17]) that will constitute the bases

\[
W_{2n}(x) = \sqrt{2} \sum h_k W_n(2x - k),
\]

\[
W_{2n+1}(x) = \sqrt{2} \sum g_k W_n(2x - k).
\]

We define

\[
m_0(\xi) = 1/\sqrt{2} \sum h_k e^{-ik\xi},
\]

\[
m_1(\xi) = -e^{i\xi} \bar{m}_0(\xi + \pi) = \frac{1}{\sqrt{2}} \sum g_k e^{ik\xi}.
\]

The well-known self-similarity relation can be written as

\[
\hat{W}_0(\xi) = m_0(\xi/2) \hat{W}_0(\xi/2),
\]

or

\[
\hat{W}_0(\xi) = \prod_{j=1}^{\infty} m_0(\xi/2^j),
\]

and

\[
\hat{W}_1(\xi) = m_1(\xi/2) \hat{W}_0(\xi/2) = m_2(\xi/4) m_0(\xi/4) \ldots
\]

and in general

\[
\hat{W}_n(\xi) = \prod_{j=1}^{\infty} m_{\epsilon_j}(\xi/2^j),
\]

where

\[
n = \sum_{j=1}^{\infty} \epsilon_j 2^{j-1}
\]

and \( \epsilon_j = 0 \) or \( 1 \).

The library of wavelet bases is the collection of functions of the form \( W_n(2^j x - k) \), where \( l, k \in \mathbb{Z} \), \( n \in \mathbb{Z} \). Although the functions are derived from their \( \xi \) domain expressions, the parameters \( l, k, n \) roughly indicate the nature of their time support and oscillations.

We state a proposition from [17].
Proposition: Any collection of indices \((l, n, k)\) such that the intervals \(2^{l}, 2^{l}(n + 1)\) form a disjoint cover of \([0, \infty)\), and \(k\) ranges over all integers, corresponds to an orthonormal basis of \(L^2(\mathbb{R})\).

It has been shown that once we find the “best basis” for our application, a fast implementation exists for determining the coefficients with respect to the basis. This is related to the tree-structured filter bank. Concisely, a complete tree structured filter bank is considered. The last level of such a decomposition yields one output sample for a finite input frame. However, in the “best basis” approach, we do not subdivide every subband until the last level. The decision of whether to subdivide is made based on a reasonable criterion according to the application. The output of this filter bank structure yields the coefficients with respect to the “best basis.” An example of a “nonwavelet” basis derived from the structure is shown in Fig. 3. This example shows a possible subdivision of the “highpass” subband, which is unlike traditional wavelet implementations.

For every possible basis in the library, we need a real valued cost, which determines the basis selection algorithm. In our case, we can look at it as a constrained minimization problem. The goal is as follows.

\[
\text{Minimize } M \text{ subject to } C_f < C. 
\]

\(C_f\) is related to the computational complexity due to the filter bank structure, and \(M\) is the cost due to the bit rate given the filter bank structure. \(C\) is the maximum limit on computations permitted. \(C_f\) is the estimated computational complexity at a particular step of the algorithm, and \(M = \text{SUPER}\) for the given filter bank structure. At every stage, a decision is made whether to decompose the subband further based on \(M\). If the decomposition results in a smaller \(M\), it is carried out. Otherwise, the decomposition is halted. At any stage, the subband that has maximum \(\text{SUPER}\) is examined first.

\(C_f\) is calculated based on the computations per frame that will result in the decomposition. In our case, the first level yields \(C_f = c\) (which was defined earlier as the computations per frame for subband decomposition), the second decision to decompose yields \(C_f = c + c/2\) (note that the number of data points decreases as the depth increases), and so on. This value keeps increasing as we decompose further. If \(C_f > C\), which is the maximum allowed computation (depending on whether the encoder or decoder imposes the more stringent constraint), the algorithm halts. In order to prevent repeated decomposition of the same frequency range, which will result in progressive loss of temporal resolution, we decompose completely within the current level before proceeding.

\(\text{SUPER}\) is the bit rate dictated by the perceptual block for transparent coding given the current subband structure. The main point to note here is that the \(\text{SUPER}\) uses the minimum noise threshold of all constituent threshold calculation partitions in order to ensure transparent encoding. Subdividing the subband might yield a smaller \(\text{SUPER}\) if one of the subbands has a tone (for instance) that increases the threshold. In that case, the subdivision occurs and, hence, the filter bank structure adapts according to the signal-dependant noise-masking threshold information. However, if a subband has a contribution to the \(\text{SUPER}\) that is smaller than a threshold \(c\), the subdivision is not carried out. This reduces the computational burden.

Another issue in this adaptive time–frequency representation is the tradeoff in resolution. As we decompose further down, we sacrifice temporal resolution for frequency resolution. The last level of decomposition has minimum temporal resolution and has the best frequency resolution. While this is a justification for the use of the perceptual output, which is calculated from an FFT, there is a drawback to sacrificing temporal resolution. The bit allocation scheme works based on some temporal correlations, and as we lose this, the bit allocation scheme is expected to perform worse. Second, there could be bursts and sudden changes in the signal, which need a good temporal resolution. Keeping these in mind, the decision on whether to decompose is carried out top-down instead of bottom-up, as it is done in other works. This way, we evaluate the signal at a better temporal resolution before we decide to decompose. The burst will typically yield a wide range of thresholds within the band, which will boost the \(\text{SUPER}\), causing the decomposition to occur until a region of high noise threshold is identified. Figs. 4 and 5 show some examples of the selection of bases. The example shows a higher threshold in the lower frequency region due to the high-energy peak in the same region. The tree structures show the resulting adaptation, where the tree subdivides further in the lower subbands, thus
lowering the bit rate (to the extent that the quantization error remains below the threshold).

Proposition: The algorithm yields the “best basis” (minimum cost) for the given computational complexity and range of temporal resolution.

An inductive proof is presented as the Appendix. This is not a globally optimal solution if we do not constrain the temporal resolution. In that case, we would need an exhaustive search to find the basis. However, in practical applications, neither do we resort to an exhaustive search, nor do we insist on the temporal resolution strictly. We instead take the middle approach and prevent the algorithm from repeatedly subdividing the same frequency range, which works for practical purposes.

V. THE BIT ALLOCATION

The bit allocation block in Fig. 1 follows the decomposition of the signal. The bit allocation proceeds with a fixed number of iterations of a zero-tree algorithm [13] before a perceptual evaluation is done. The algorithm organizes the coefficients in a tree structure that is temporally aligned from coarse to fine. It classifies the coefficients into one of four types: POS, NEG, IZ, and ZTR. At each iteration, we have a threshold, which is half the maximum amplitude to begin with. POS refers to positive, with amplitude greater than the threshold, and NEG refers to negative, with amplitude greater than the threshold. ZTR means that this coefficient and all its descendants are below the threshold, and IZ means that this coefficient is below threshold but that there is a descendant that is above the threshold. The decoder reconstructs as a 1 the first bit for POS and NEG coefficients and as a zero for a ZTR coefficient and all its descendants. The IZ coefficients are also reconstructed as 0. The next pass refines the POS and NEG coefficients further by supplying another bit, thus reducing the level of uncertainty by half. The threshold is reduced by half, and the next iteration proceeds.

The perceptual evaluation step between two iterations checks if the bit rate, as dictated by the thresholds, is satisfied. If it is, the algorithm terminates.

This algorithm tries to exploit the remnants of temporal correlations that exist in the wavelet packet coefficients. The complete data frame now comprises the header and the bit stream. The header consists of the following.

- The filter bank structure as described in the previous section is judiciously encoded by exploiting the tree struc-
ture. Starting from the root of the tree, a “1” indicates that the subband is further decomposed, and a “0” indicates that it is not further decomposed (which the decoder interprets as having no descendants).

- We have the number of iterations of the zero tree and the classification information.

This is followed by the bit stream. Following this, the entire frame is compressed by lossless coding and transmitted.

VI. THE COMPLETE ALGORITHM

As shown in Fig. 1, the implementation of the algorithm involves the filter bank structure, the psychoacoustic model, and the bit allocation scheme. All results presented in this paper use the filter banks that implement the spline-based biorthogonal wavelet transform [21]. In this structure, the analysis filters are different from the synthesis filters. The advantage of using the biorthogonal filter banks is that it allows us to experiment with various orders of piecewise polynomials as the bases functions. The filters are FIR, yield perfect reconstruction, and cancel aliasing. The implementation of the psychoacoustic model essentially follows the MPEG specification; the bit allocation scheme follows the zero-tree algorithm.

The complete algorithm incorporates the process of wavelet packet decomposition and bit allocation in an iterative scheme. Frames of input audio are used as input to two blocks: the perceptual block and the filter bank structure. The perceptual block calculates the masking threshold for the critical bands as described in [14] and spreads it over the frequency lines. This enables calculation of the SUPER cost value that determines the choice of the tree structure. The allowed complexity of the decoder/encoder also determines the depth of the decomposition. A running variable accumulates the computational complexity incurred due to each additional filtering operation, considering that the number of samples that are filtered is halved at each subdivision. When this variable equals the limit on the number of computations per frame, the filter banks are not allowed to subdivide further. The allowed noise as dictated by the psychoacoustic model, and the bit budget determines the stopping criteria of the zero-tree bit allocation. Each iteration of the zero-tree allocation scheme is followed by a perceptual evaluation step to determine a stopping rule. This concept of bit allocation using the bit budget as the limit, followed by perceptual evaluation, is akin to the two-loop encoding scheme of MPEG Layer III. Although the zero-tree algorithm is used in this work, suitable modifications can be made to adapt the algorithm to recent improvements to the zero-tree method, e.g., [22], that are based on ordering coefficients by magnitude and set partitioning. Further reductions in bit rates could be achieved by incorporating such improved methods.

The decomposition and bit allocation iterations for the technique described in this paper are outlined in Fig. 6 by a flow chart for the encoder.

VII. EXPERIMENTS AND RESULTS

The experiments performed in this work are based on perceptual evaluation since a mean squared error measure is inappropriate in this case. This is because our filter bank structure used a perceptual cost criterion rather than one based on mean squared error. In order to extract samples for the experimental data, we extracted data from digitally recorded CD’s. The digital data from the CD player was gathered for a duration of 10–15 s. These samples were chosen from a variety of instruments. Western classical music was chosen for samples of violin and flute in isolation. Orchestral-type music was chosen from saxophone and sitar recitals of Indian classical music. A combination of vocal and orchestral music was chosen in a segment of Indian popular film music. Attention was chosen to selecting segments that have particularly strong attacks in order to evaluate the performance of the compression scheme. The source data was 44.1 kHz, 16 bit PCM samples.

The ten subjects were chosen from a pool of graduate students primarily from Electrical and Computer Engineering. Most of the students are not experienced in listening to distortion resulting from compression algorithms. The students were presented a pair of samples that were separated by a pause of 5 s. At the end of the two segments, the subject
TABLE I
RESULTS OF PERCEPTUAL TESTS ON WAVELET PACKET BASED PERCEPTUAL COMPRESSION

<table>
<thead>
<tr>
<th>Music Type</th>
<th>Likelihood of Listener Preferring the Original over Reconstructed</th>
</tr>
</thead>
<tbody>
<tr>
<td>violin</td>
<td>0.5</td>
</tr>
<tr>
<td>violin and viola</td>
<td>0.4962</td>
</tr>
<tr>
<td>flute</td>
<td>0.5142</td>
</tr>
<tr>
<td>sitar</td>
<td>0.5</td>
</tr>
<tr>
<td>film tune (vocal and orchestra)</td>
<td>0.5</td>
</tr>
<tr>
<td>saxophone</td>
<td>0.9072</td>
</tr>
</tbody>
</table>

was asked to enter the scores and proceed by pressing a key. The subjects were told that one or both or neither of the samples could be the result of reconstruction from a compression algorithm. They had to identify distorted samples and were allowed to give a “both are good” answer. All experiments were performed on monophonic samples. This experimental methodology follows [12], and the results are also reported in a similar manner. Table I shows, for each piece of music, the likelihood that the original was preferred over the reconstructed. A score of 0.5 means that an equal number of listeners preferred the original as preferred the reconstructed.

We also measure the quality on a scale of 0 to 5 as the order of the spline wavelet was increased. It was noted that as the order increased beyond degree 5, there was no perceivable advantage to quality. Although it is true that the higher degree polynomials yield a smoother reconstruction, it does not seem to reflect immediately on the quality of the output. Spline order 5 reconstruction was considered superior to order 3 by all listeners, but a further increase did not yield better quality.

“C” libraries and examples pertaining to this implementation can be downloaded from the internet site http://wavelet.ecn.purdue.edu/~speech/.

VIII. CONCLUSIONS

In this paper, we have constructed a wavelet packet-based compression scheme suitable for high-quality audio transfers over the internet or storage at about 45 kb/s. It relies on an adaptive structure where the encoder can adapt to a given computational complexity while maintaining transparency. It can also respond to a changing bit budget and is amenable to progressive transmission, where the best reconstruction is possible given the file size. It brings together several powerful compression techniques including subband coding, perceptual noise masking, and bit allocation using the zero tree method.

The results in Table I show that for most of the samples tested, the likelihood of the listener identifying the reconstruction from our algorithm is close to 0.5, which implies transparent quality. The flute is an instance where the algorithm did not perform as well as it did for the other examples. Further work is necessary to study the problem of compressing wind instruments by this algorithm.

Although the perceptual tests were not extensive, our results so far show that this is a promising method of high-quality audio compression. It has a simple decoder structure, and the bit stream has several desirable properties. The computational adaptation of the algorithm and the progressive transmission property make our scheme very suitable for internet applications.

APPENDIX

PROOF OF OPTIMALITY OF THE ALGORITHM

We need to show that for a given computational complexity, the partition of the frequency spectrum yielded by the algorithm results in the subband decomposition where the bits needed for transparent encoding is the least (considering only partitions that the dyadic tree gives rise to), given the range of temporal resolution. For the first step—\( k = 0 \), given \( c < C \)—the obvious partition is the subdivision into two equal divisions: lowpass and highpass. This reduces \( M \) in all but the pathological case, where the minimum thresholds are identical in both partitions. Hence, given \( C = c \), this is the best partition.

Given a partition at stage \( k \), choose the subband that causes the best reduction to the cost, provided the total computation \( c_k \) does not exceed \( C \). The SUPER is forced to strictly decrease whenever a decomposition is made. Otherwise, the decomposition is not carried out. When the decomposition is made, if the subband before decomposition was \( S \) and the parts after division are \( S' \) and \( S'' \), the decomposition occurs only when

\[
-\log(t_s/c_s) < -\log(t_s'/c_s') - \log(t_s''/c_s'').
\]

That is, the sum of the thresholds (normalized by the energies in the subbands) is increased. This is due to the fact that the threshold chosen is the minimum of the constituent critical bands, as explained previously. This means that the quantization error permitted increases; hence, SUPER decreases. Therefore, it reduces the cost as the subdivision proceeds.

Given the current computation \( c_{k+1} \), we thus have the choice that achieves the best cost reduction. Hence, by induction, when the algorithm terminates with the computations exceeding \( C \) (or just before exceeding \( C \)), we have the optimal partition or the best possible subband structure that has least SUPER. We note here that we decompose completely up to a certain level of refinement as dictated by the need to preserve temporal resolution.

Hence, under the computational constraint, no other search would yield a lower total cost for the final best basis. At the worst case, the PE will be attained when a complete decomposition is achieved at the computational expense of a full tree decomposition.

ACKNOWLEDGMENT

The authors gratefully acknowledge Prof. G. Long of the Department of Physics and the Department of Audiology and Speech Sciences, Purdue University, for having helped with the data collection and administration of the perceptual experiments. They also acknowledge the individuals who kindly participated in the experiments.
REFERENCES


Pramila Srinivasan (M’97) was born in New York, NY, in 1968. She received the B.E. degree from P.S.G. College of Technology, Coimbatore, India, in 1990 and the M.S. degree in electrical engineering from Rutgers University, New Brunswick, NJ, in 1993. She received the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, in 1997.

After receiving the Ph.D. degree, she worked at Silicon Graphics, Inc., Mountain View, CA. She is presently with Stream Machine Company, Santa Clara, CA, where she is involved in applications of audio signal processing in consumer products. Her research interests include audio and image compression, filter banks and wavelets, and various aspects of audio signal processing.

Dr. Srinivasan received the Intel Foundation Graduate Fellowship in 1995.

Leah H. Jamieson (F’93) received the S.B. degree in mathematics from the Massachusetts Institute of Technology, Cambridge, in 1972 and the M.A. and M.S.E. degrees in 1974 and the Ph.D. degree in 1977, all from the Department of Electrical Engineering and Computer Science, Princeton University, Princeton, NJ.

In 1976, she joined the faculty at Purdue University, West Lafayette, IN, where she is currently Professor of electrical and computer engineering. Her research interests include speech processing; the design and analysis of parallel processing algorithms; and the application of parallel processing to the areas of digital speech, image, and signal processing. She has co-edited the books Algorithmically Specialized Parallel Computers (New York: Academic, 1985) and The Characteristics of Parallel Algorithms (Cambridge, MA: MIT Press, 1987).

Dr. Jamieson has served as an Associate Editor for the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING and the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS. She is currently Co-Chair of the Computing Research Association’s Committee on the Status of Women in Computing Research and President of the IEEE Signal Processing Society. She has been an IEEE Signal Processing Society Distinguished Lecturer and an IEEE Computer Society Distinguished Visitor. She is Co-Founder and Co-Director of the Engineering Projects in Community Service (EPICS) Center at Purdue University and received the 1997 ASEE Chester F. Carlson Award for Innovation in Engineering Education for her work on the EPICS program.