A Review of Algorithms for Perceptual Coding of Digital Audio Signals †

Ted Painter, *Student Member IEEE*, and Andreas Spanias, *Senior Member IEEE*

Department of Electrical Engineering, Telecommunications Research Center
Arizona State University, Tempe, Arizona 85287-7206
spanias@asu.edu, painter@asu.edu

ABSTRACT

During the last decade, CD-quality digital audio has essentially replaced analog audio. During this same period, new digital audio applications have emerged for network, wireless, and multimedia computing systems which face such constraints as reduced channel bandwidth, limited storage capacity, and low cost. These new applications have created a demand for high-quality digital audio delivery at low bit rates. In response to this need, considerable research has been devoted to the development of algorithms for perceptually transparent coding of high-fidelity (CD-quality) digital audio. As a result, many algorithms have been proposed, and several have now become international and/or commercial product standards. This paper reviews algorithms for perceptually transparent coding of CD-quality digital audio, including both research and standardization activities. The paper is organized as follows. First, psychoacoustic principles are described with the MPEG psychoacoustic signal analysis model 1 discussed in some detail. Then, we review methodologies which achieve perceptually transparent coding of FM- and CD-quality audio signals, including algorithms which manipulate transform components and subband signal decompositions. The discussion concentrates on architectures and applications of those techniques which utilize psychoacoustic models to exploit efficiently masking characteristics of the human receiver. Several algorithms which have become international and/or commercial standards are also presented, including the ISO/MPEG family and the Dolby AC-3 algorithms. The paper concludes with a brief discussion of future research directions.

I. INTRODUCTION

Audio coding or *audio compression* algorithms are used to obtain compact digital representations of high-fidelity (wideband) audio signals for the purpose of efficient transmission or storage. The central objective in audio coding is to represent the signal with a minimum number of bits while achieving transparent signal reproduction, i.e., while generating output audio which cannot be distinguished from the original input, even by a sensitive listener ("golden ears"). This paper gives a review of algorithms for transparent coding of high-fidelity audio.

The introduction of the compact disk (CD) in the early eighties [1] brought to the fore all of the advantages of digital audio representation, including unprecedented high-fidelity, dynamic range, and robustness. These advantages, however, came at the expense of high data rates. Conventional CD and digital audio tape (DAT) systems are typically sampled at 44.1 or 48 kilohertz (kHz), using pulse code modulation (PCM) with a sixteen bit sample resolution. This results in uncompressed data rates of 705.6/768 kilobits per second (kbps) for a monaural channel, or 1.41/1.54 megabits per second (Mbps) for a stereo pair at 44.1/48 kHz, respectively. Although high, these data rates were accommodated successfully in first generation digital audio applications such as CD and DAT. Unfortunately, second generation multimedia applications and wireless systems in particular are often subject to bandwidth or cost constraints which are incompatible with high data rates. Because of the success enjoyed by the first generation, however, end users have come to expect "CD-quality" audio reproduction from any digital system. New network and wireless multimedia digital audio systems, therefore, must reduce data rates without compromising reproduction quality. These and other considerations have motivated considerable research during the last decade towards formulation of compression schemes which can satisfy simultaneously the conflicting demands of high compression ratios and transparent reproduction quality for high-fidelity audio signals [2][3][4][5][6][7][8][9][10][11]. As a result, several standards have been developed [12][13][14][15], particularly in the last five years [16][17][18][19], and several are now being deployed commercially [94] [97][100][102] (Table 2).

A. GENERIC PERCEPTUAL AUDIO CODING ARCHITECTURE

This review considers several classes of analysis-synthesis data compression algorithms, including those

† Portions of this work have been sponsored by a grant from the NDTC Committee of the Intel Corporation. Direct all communications to A. Spanias.

which manipulate: transform components, time-domain sequences from critically sampled banks of bandpass filters, linear predictive coding (LPC) model parameters, or some hybrid parametric set. We note here that although the enormous capacity of new storage media such as Digital Versatile Disc (DVD) can accommodate *lossless* audio coding [20][21], the research interest and hence all of the algorithms we describe are *lossy* compression schemes which seek to exploit the psychoacoustic principles described in section two. Lossy schemes offer the advantage of lower bit rates (e.g., less than 1 bit per sample) relative to lossless schemes (e.g., 10 bits per sample). Naturally, there is a debate over the quality limitations associated with lossy compression. In fact, some experts believe that *uncompressed* digital CD-quality audio (44.1 kHz/16b) is intrinsically inferior to the analog original. They contend that sample rates above 55 kHz and word lengths greater than 20 bits [21] are necessary to achieve transparency in the absence of any compression. It is beyond the scope of this review to address this debate.

Before considering different classes of audio coding algorithms, it is first useful to note the architectural similarities which characterize most perceptual audio coders. The lossy compression systems described throughout the remainder of this review achieve coding gain by exploiting both *perceptual irrelevancies* and *statistical redundancies*. All of these algorithms are based on the generic architecture shown in Fig. 1. The coders typically segment input signals into quasi-stationary frames ranging from 2 to 50 milliseconds in duration. A time-frequency analysis section then decomposes each analysis frame. The time/frequency analysis approximates the temporal and spectral analysis properties of the human auditory system. It transforms input audio into a set of parameters which can be quantized and encoded according to a perceptual distortion metric. Depending on overall system objectives and design philosophy, the time-frequency analysis section might contain a

- Unitary transform
- Time-invariant bank of uniform bandpass filters
- Time-varying (signal-adaptive), critically sampled bank of non-uniform bandpass filters
- Hybrid transform/filterbank signal analyzer
- Harmonic/sinusoidal analyzer
- Source-system analysis (LPC/Multipulse excitation)

The choice of time-frequency analysis methodology always involves a fundamental tradeoff between time and frequency resolution requirements. Perceptual distortion control is achieved by a psychoacoustic signal analysis section which estimates signal masking power based on psychoacoustic principles (see section two). The psychoacoustic model delivers masking thresholds which quantify the maximum amount of distortion that can be injected at each point in the time-frequency plane during quantization and encoding of the time-frequency parameters without introducing audible artifacts in the reconstructed signal. The psychoacoustic model therefore allows the quantization and encoding section to exploit perceptual irrelevancies in the time-frequency parameter set. The quantization and encoding section can also exploit statistical redundancies through classical techniques such as differential pulse code modulation (DPCM) or adaptive DPCM (ADPCM). Quantization might be uniform or pdf-optimized (Lloyd-Max), and it might be performed on either scalar or vector quantities (VQ). Once a quantized compact parametric set has been formed, remaining redundancies are typically removed through run-length (RL) and entropy (e.g. Huffman, arithmetic, LZW) coding techniques. Since the psychoacoustic distortion control model is signal adaptive, most algorithms are inherently variable rate. Fixed channel rate requirements are usually satisfied through buffer feedback schemes, which often introduce encoding delays.

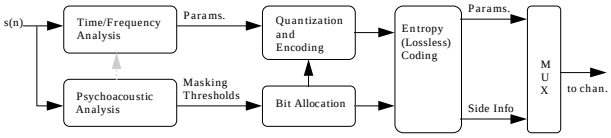


Fig. 1. Generic Perceptual Audio Encoder

The study of perceptual entropy (PE) suggests that transparent coding is possible in the neighborhood of 2 bits per sample [45] for most for high-fidelity audio sources (~88 kpbs given 44.1 kHz sampling). The lossy perceptual coding algorithms discussed in the remainder of this paper confirm this possibility. In fact, several coders approach transparency in the neighborhood of 1 bit per sample. Regardless of design details, all perceptual audio coders seek to achieve transparent quality at low bit rates with tractable complexity and manageable delay. The discussion of algorithms given in sections three through five brings to light many of the tradeoffs involved with the various coder design philosophies.

B. PAPER ORGANIZATION

The rest of the paper is organized as follows. In section II, psychoacoustic principles are described which can be exploited for significant coding gain. Johnston's notion of perceptual entropy is presented as a measure of the fundamental limit of transparent compression for audio. Sections III through V review state-of-the-art algorithms which achieve transparent coding of FM- and CD-quality audio signals, including several techniques which are established in international standards. Transform coding methodologies are described in section III, and subband coding algorithms are addressed in section IV. In addition to methods based on uniform bandwidth filterbanks, section IV covers coding methods which utilize discrete wavelet transforms

and non-uniform filterbanks. Finally, section V is concerned with standardization activities in audio coding. It describes recently adopted standards including the ISO/IEC MPEG family, the Phillips' Digital Compact Cassette (DCC), the Sony Minidisk, and the Dolby AC-3 algorithms. The paper concludes with a brief discussion of future research directions.

For additional information, one can also refer to informative reviews of recent progress in wideband and hi-fidelity audio coding which have appeared in the literature. Discussions of audio signal characteristics and the application of psychoacoustic principles to audio coding can be found in [22],[23], and [24]. Jayant, *et al.* of Bell Labs also considered perceptual models and their applications to speech, video, and audio signal compression [25]. Noll describes current algorithms in [26] and [27], including the ISO/MPEG audio compression standard.

II. PSYCHOACOUSTIC PRINCIPLES

High precision engineering models for high-fidelity audio currently do not exist. Therefore, audio coding algorithms must rely upon generalized receiver models to optimize coding efficiency. In the case of audio, the receiver is ultimately the human ear and sound perception is affected by its masking properties. The field of psychoacoustics [28][29][30][31][32][33][34] has made significant progress toward characterizing human auditory perception and particularly the time-frequency analysis capabilities of the inner ear. Although applying perceptual rules to signal coding is not a new idea [35], most current audio coders achieve compression by exploiting the fact that "irrelevant" signal information is not detectable by even a well trained or sensitive listener. Irrelevant information is identified during signal analysis by incorporating into the coder several psychoacoustic principles, including absolute hearing thresholds, critical band frequency analysis, simultaneous masking, the spread of masking along the basilar membrane, and temporal masking. Combining these psychoacoustic notions with basic properties of signal quantization has also led to the development of perceptual entropy [36], a quantitative estimate of the fundamental limit of transparent audio signal compression. This section reviews psychoacoustic fundamentals and perceptual entropy, then gives as an application example some details of the ISO/MPEG psychoacoustic model one.

A. ABSOLUTE THRESHOLD OF HEARING

The absolute threshold of hearing is characterized by the amount of energy needed in a pure tone such that it can be detected by a listener in a noiseless environment. The frequency dependence of this threshold was quantified as early as 1940, when Fletcher [28] reported test results for a range of listeners which were generated in an NIH study of typical American hearing acuity. The

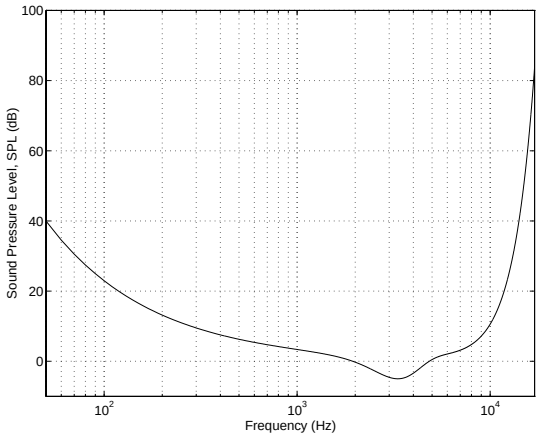


Fig. 2. The Absolute Threshold of Hearing

quiet threshold is well approximated [37] by the non-linear function

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 \quad \text{(dB SPL)} \qquad (1)$$

which is representative of a young listener with acute hearing. When applied to signal compression, $T_q(f)$ can be interpreted as a maximum allowable energy level for coding distortions introduced in the frequency domain (Fig. 2). Algorithm designers have no *a priori* knowledge regarding actual playback levels, therefore the sound pressure level (SPL) curve is often referenced to the coding system by equating the lowest point on the curve (i.e., 4 kHz) to the energy in +/- 1 bit of signal amplitude. Such a practice is common in algorithms which utilize the absolute threshold of hearing.

B. CRITICAL BANDS

Using the absolute threshold of hearing to shape the coding distortion spectrum represents the first step towards perceptual coding. Next we consider how the ear actually does spectral analysis. It turns out that a frequency-to-place transformation takes place in the inner ear, along the basilar membrane. Distinct regions in the cochlea, each with a set of neural receptors, are "tuned" to different frequency bands. Empirical work by several observers led to the modern notion of critical bands [28][29][30][31] which correspond to these cochlear regions. In the experimental sense, critical bandwidth can be loosely defined as the bandwidth at which subjective responses change abruptly. For example, the perceived loudness of a narrowband noise source at constant sound pressure level remains constant even as the bandwidth is increased up to the critical bandwidth. The loudness then begins to increase. In a different experiment (Fig 3a), the detection threshold for a narrowband noise source between two masking tones remains constant as long as the frequency separation between the tones remains within a critical bandwidth. Beyond

this bandwidth, the threshold rapidly decreases (Fig 3c).

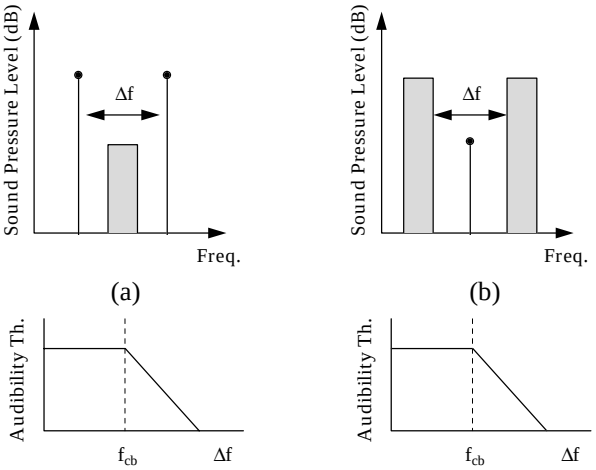


Fig. 3. Critical Band Measurement Methods

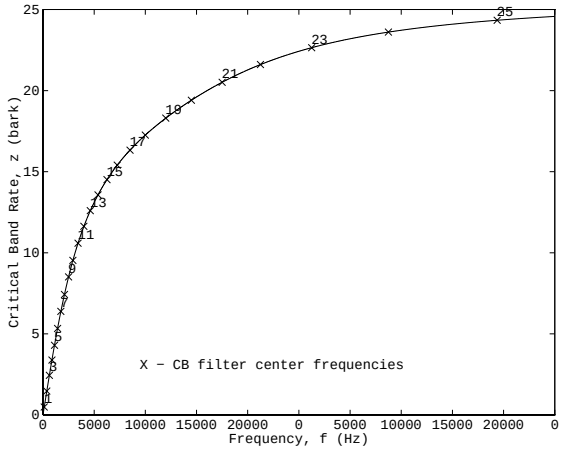


Fig. 4. (a) Critical Band Rate, *z(f)*, and (b) Critical Bandwidth, BW_c

A similar notched-noise experiment can be constructed with masker and maskee roles reversed (Fig. 3b,d). Critical bandwidth tends to remain constant (about 100 Hz) up to 500 Hz, and increases to approximately 20% of the center frequency above 500 Hz. For an average listener, critical bandwidth (Fig. 4b) is conveniently approximated [33] by

$$BW_c(f) = 25 + 75\left[1 + 1.4(f/1000)^2\right]^{0.69} \text{ (Hz)} \quad (2)$$

Although the function BW_c is continuous, it is useful when building practical systems to treat the ear as a discrete set of bandpass filters which obeys Eq. (2). Table 1 gives an idealized filterbank which corresponds to the discrete points labeled on the curve in Figs. 4a, 4b. A distance of 1 critical band is commonly referred to as "one bark" in the literature. The function [33]

$$z(f) = 13\arctan(.00076f) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right] \text{ (Bark)} \quad (3)$$

is often used to convert from frequency in Hertz to the bark scale (Fig 4a). Corresponding to the center frequencies of the Table 1 filterbank, the numbered points in Fig. 4a illustrate that the non-uniform Hertz spacing of the filterbank (Fig. 5) is actually uniform on a bark scale. Thus, one critical bandwidth comprises one bark. Intra-band and inter-band masking properties associated with the ear's critical band mechanisms are routinely used by modern audio coders to shape the coding distortion spectrum. These masking properties are described next.

Band No.	Center Freq. (Hz)	Bandwidth (Hz)
1	50	-100
2	150	100-200
3	250	200-300
4	350	300-400
5	450	400-510
6	570	510-630
7	700	630-770
8	840	770-920
9	1000	920-1080
11	1370	1270-1480
12	1600	1480-1720
13	1850	1720-2000
14	2150	2000-2320
15	2500	2320-2700
16	2900	2700-3150
17	3400	3150-3700
18	4000	3700-4400
19	4800	4400-5300
20	5800	5300-6400
21	7000	6400-7700
22	8500	7700-9500
23	10,500	9500-12000
24	13,500	12000-15500
25	19,500	15500-

Table 1 Critical Band Filterbank [after Scharf]

C. SIMULTANEOUS MASKING AND THE SPREAD OF MASKING

Masking refers to a process where one sound is rendered inaudible because of the presence of another sound. Simultaneous masking refers to a frequency-

domain phenomenon which has been observed within critical bands (in-band). For the purposes of shaping coding distortions it is convenient to distinguish between two types of simultaneous masking, namely *tone-masking-noise* [31], and *noise-masking-tone* [32]. In the first case, a tone occurring at the center of a critical band masks noise of any subcritical bandwidth or shape,

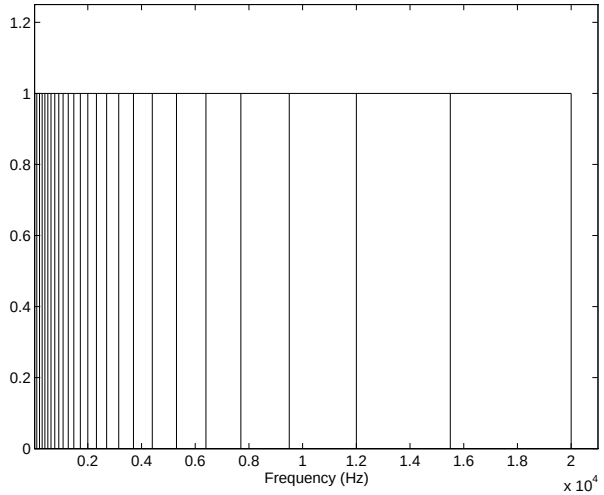


Fig. 5. Idealized Critical Band Filterbank

provided the noise spectrum is below a predictable threshold directly related to the strength of the masking tone. The second masking type follows the same pattern with the roles of masker and maskee reversed. A simplified explanation of the mechanism underlying both masking phenomena is as follows. The presence of a strong noise or tone masker creates an excitation of sufficient strength on the basilar membrane at the critical band location to effectively block transmission of a weaker signal. Inter-band masking has also been observed, i.e., a masker centered within one critical band has some predictable effect on detection thresholds in other critical bands. This effect, also known as the spread of masking, is often modeled in coding applications by an approximately triangular spreading function which has slopes of +25 and -10 dB per bark. A convenient analytical expression [35] is given by:

$$SF_{dB}(x) = 15.81 + 7.5(x + 0.474) - 17.5\sqrt{1 + (x + 0.474)^2}\ \text{dB} \qquad (4)$$

where x has units of barks and $SF_{db}(x)$ is expressed in dB. After critical band analysis is done and the spread of masking has been accounted for, masking thresholds in psychoacoustic coders are often established by the [38] decibel (dB) relations:

$$TH_N = E_T - 14.5 - B \qquad (5)$$

$$TH_T = E_N - K \qquad (6)$$

where TH_N and TH_T, respectively, are the noise and tone masking thresholds due to tone-masking noise and noise-masking-tone, E_N and E_T are the critical band noise and tone masker energy levels, and B is the critical band number. Depending upon the algorithm, the parameter K has typically been set between 3 and 5 dB. Masking thresholds are commonly referred to in the literature as (bark scale) functions of just noticeable distortion (JND). One psychoacoustic coding scenario might involve first classifying masking signals as either noise or tone, next computing appropriate thresholds, then using this information to shape the noise spectrum beneath JND. Note that the absolute threshold (T_{ABS}) of

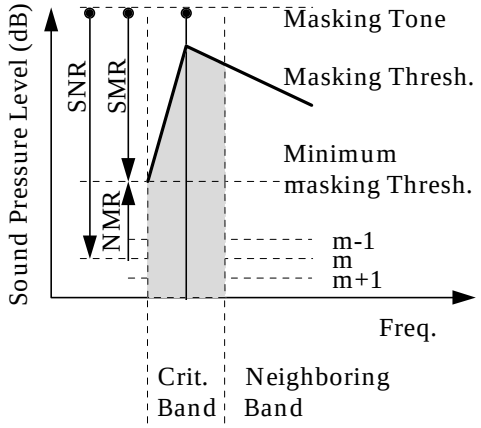


Fig. 6. Schematic Representation of Simultaneous Masking (after [26])

hearing is also considered when shaping the noise spectra, and that MAX(JND, T_{ABS}) is most often used as the permissible distortion threshold. Notions of critical bandwidth and simultaneous masking in the audio coding context give rise to some convenient terminology illustrated in Fig. 6, where we consider the case of a single masking tone occurring at the center of a critical band. All levels in the figure are given in terms of dB SPL. A hypothetical masking tone occurs at some masking level. This generates an excitation along the basilar membrane which is modeled by a spreading function and a corresponding *masking threshold*. For the band under consideration, the *minimum masking threshold* denotes the spreading function in-band minimum. Assuming the masker is quantized using an m-bit uniform scalar quantizer, noise might be introduced at the level m. *Signal-to-mask ratio* (SMR) and noise-to-mask ratio (NMR) denote the log distances from the minimum masking threshold to the masker and noise levels, respectively.

D. TEMPORAL MASKING

Masking also occurs in the time-domain. In the context of audio signal analysis, abrupt signal transients (e.g., the onset of a percussive musical instrument) create pre- and post- masking regions in time during which a listener will not perceive signals beneath the elevated audibility thresholds produced by a masker. The skirts on both regions are schematically represented in Fig. 7. In other words, absolute audibility thresholds for masked sounds are artificially increased prior to, during, and following the occurrence of a masking signal. Whereas premasking tends to last only about 5 ms,

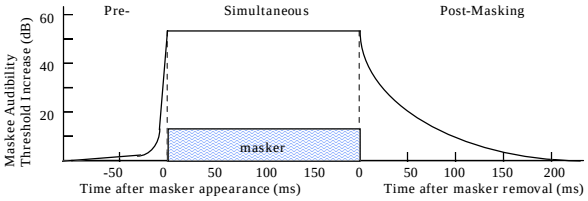


Fig. 7. Schematic Representation of Temporal Masking Properties of the Human Ear (after [33])

postmasking will extend anywhere from 50 to 300 ms, depending upon the strength and duration of the masker [33][39]. Temporal masking has been used in several audio coding algorithms. Pre-masking in particular has been exploited in conjunction with adaptive block size transform coding to compensate for pre-echo distortions (section III).

E. PERCEPTUAL ENTROPY

Johnston at Bell Labs has combined notions of psychoacoustic masking with signal quantization principles to define perceptual entropy (PE), a measure of perceptually relevant information contained in any audio record. Expressed in bits per sample, PE represents a theoretical limit on the compressibility of a particular signal. PE measurements reported in [36] and [6] suggest that a wide variety of CD quality audio source material can be transparently compressed at approximately 2.1 bits per sample. The PE estimation process is accomplished as follows. The signal is first windowed and transformed to the frequency domain. A masking threshold is then obtained using perceptual rules. Finally, a determination is made of the number of bits required to quantize the spectrum without injecting perceptible noise. The PE measurement is obtained by constructing a PE histogram over many frames and then choosing a worst-case value as the actual measurement.

The frequency-domain transformation is done with a Hanning window followed by a 2048-point FFT. Masking thresholds are obtained by performing critical band analysis (with spreading), making a determination of the noiselike or tonelike nature of the signal, applying thresholding rules for the signal quality, then accounting for the absolute hearing threshold. First, real and imaginary transform components are converted to power spectral components

$$P(\omega) = \mathrm{Re}^2(\omega) + \mathrm{Im}^2(\omega) \tag{7}$$

then a discrete bark spectrum is formed by summing the energy in each critical band (Table 1)

$$B_i = \sum_{\omega=bl_i}^{bh_i} P(\omega) \tag{8}$$

where the summation limits are the critical band boundaries. The range of the index, i, is sample rate dependent, and in particular $i \in \{1,25\}$ for CD-quality signals. A basilar spreading function (Eq.4) is then convolved with the discrete bark spectrum

$$C_i = B_i * SF_i \tag{9}$$

to account for inter-band masking. An estimation of the tonelike or noiselike quality for C_i is then obtained using the spectral flatness measure [40] (SFM)

$$SFM = \frac{\mu_g}{\mu_a} \tag{10}$$

where μ_g and μ_a correspond to the geometric and arithmetic means of the PSD components for each band. The SFM has the property that it is bounded by 0 and 1. Values close to 1 will occur if the spectrum is flat in a particular band, indicating a decorrelated (noisy) band. Values close to zero will occur if the spectrum in a particular band is nearly sinusoidal. A "coefficient of tonality," α, is next derived from the SFM on a dB scale

$$\alpha = \min\left(\frac{SFM_{dB}}{-60}, 1\right) \tag{11}$$

and this is used to weight the thresholding rules given by Eq. (5) and Eq. (6) [with K = 5.5] as follows for each band to form an offset

$$O_i = \alpha(14.5 + i) + (1 - \alpha)5.5 \text{ (in dB)} \tag{12}$$

A set of JND estimates in the frequency power domain are then formed by subtracting the offsets from the bark spectral components

$$T_i = 10^{\log_{10}(C_i) - \frac{O_i}{10}} \tag{13}$$

These estimates are scaled by a correction factor to simulate deconvolution of the spreading function, then each T_i is checked against the absolute threshold of hearing and replaced by $\max\left(T_i, T_{ABS}(i)\right)$. As previously noted, the absolute threshold is referenced to the energy in a 4 kHz sinusoid of +/- 1 bit amplitude. By applying uniform quantization principles to the signal and associated set of JND estimates, it is possible to estimate a lower bound on the number of bits required to achieve transparent coding. In fact, it can be shown that the perceptual entropy in bits per sample is given by

$$PE = \sum_{i=1}^{25} \sum_{\omega=bl_i}^{bh_i} \log_2\left(2\left|n\,\mathrm{int}\left(\frac{\mathrm{Re}(\omega)}{\sqrt{6T_i/k_i}}\right)\right| + 1\right) + \log_2\left(2\left|n\,\mathrm{int}\left(\frac{\mathrm{Im}(\omega)}{\sqrt{6T_i/k_i}}\right)\right| + 1\right) \text{ (bits/sample)} \tag{14}$$

where i is the index of critical band, bl_i and bh_i are the upper and lower bounds of band i, k_i is the number of transform components in band i, T_i is the masking threshold in band i (Eq. (13)), and *nint* denotes rounding to the nearest integer. Note that if 0 occurs in the log we assign 0 for the result.

The masking thresholds used in the above PE computation also form the basis for a transform coding algorithm described in section III.

F. PRE-ECHO DISTORTION

A problem known as "pre-echo" can arise in transform coders using perceptual coding rules. Pre-echoes occur when a signal with a sharp attack begins near the end of a transform block immediately following a region of low energy. This situation can arise when coding recordings of percussive instruments such as the castanets, for example (Fig 8a). The inverse transform spreads quantization distortion evenly throughout the reconstructed block according to the relatively lax masking thresholds associated with the block average spectral estimate (Fig 8b), resulting in unmasked distortion in the low energy region preceding in time the signal attack at the decoder. Although it has the potential to compensate for pre-echo, temporal premasking is possible only if the transform block size is sufficiently small (minimal coder delay). A more robust solution to the problem relies upon the use of adaptive transform block sizes. Long blocks are applied during steady-state audio segments, and short blocks are applied when pre-echo is likely. Several algorithms make use of this approach.

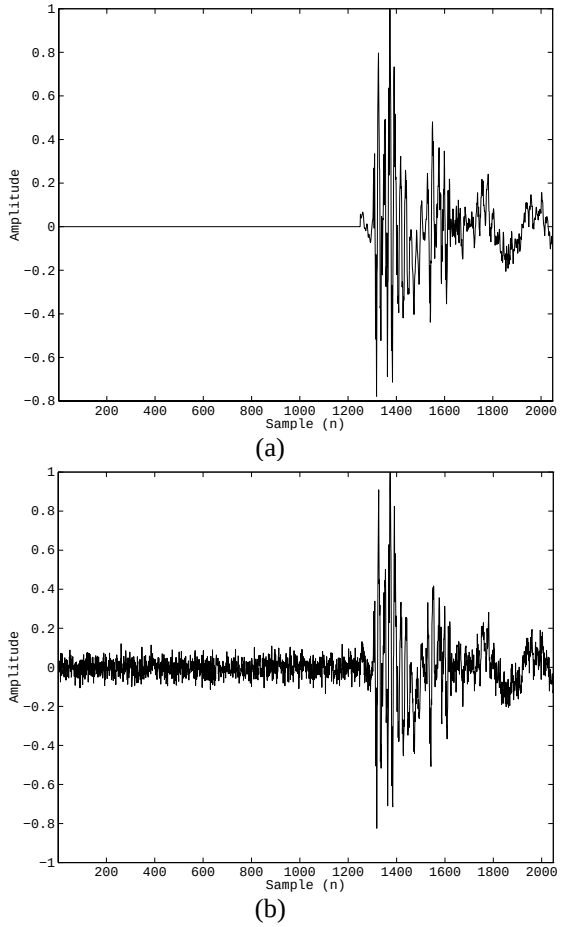


Fig. 8. Pre-Echo Example: (a) Uncoded Castanets. (b) Transform Coded Castanets, 2048-Point Block Size

G. APPLICATION OF PSYCHOACOUSTIC PRINCIPLES: ISO 11172-3 (MPEG-1) PSYCHOACOUSTIC MODEL 1

It is useful to consider an example of how the psychoacoustic principles described thus far are applied in actual coding algorithms. The ISO/IEC 11172-3 (MPEG-1, layer 1) psychoacoustic model 1 [17] determines the maximum allowable quantization noise energy in each critical band such that quantization noise remains inaudible. In one of its modes, the model uses a 512-point DFT for high resolution spectral analysis (86.13 Hz), then estimates for each input frame individual simultaneous masking thresholds due to the presence of tone-like and noise-like maskers in the signal spectrum. A global masking threshold is then estimated for a subset of the original 256 frequency bins by (power) additive combination of the tonal and non-tonal individual masking thresholds. The remainder of this section describes the step-by-step model operations. Sample results are given for one frame of CD-quality pop music sampled at 44.1 kHz/16-bits per sample. The five steps leading to computation of global masking thresholds are as follows:

STEP 1: SPECTRAL ANALYSIS AND SPL NORMALIZATION

First, incoming audio samples, $s(n)$, are normalized according to the FFT length, N, and the number of bits per sample, b, using the relation

$$x(n) = \frac{s(n)}{N\left(2^{b-1}\right)} \tag{15}$$

Normalization references the power spectrum to a 0-dB maximum. The normalized input, $x(n)$, is then segmented into 12 ms frames (512 samples) using a 1/16th-overlapped Hann window such that each frame contains 10.9 ms of new data. A power spectral density (PSD) estimate, $P(k)$, is then obtained using a 512-point FFT, i.e.,

$$P(k) = PN + 10 \cdot \log_{10}\left|\sum_{n=0}^{N-1} w(n)x(n)e^{-j\frac{2\pi kn}{N}}\right|^2 \quad 0 \le k \le \frac{N}{2} \tag{16}$$

where the power normalization term, PN, is fixed at 90 dB and the Hann window, $w(n)$, is defined as

$$w(n) = \frac{1}{2}\left[1 - \cos\left(\frac{2\pi n}{N}\right)\right] \tag{17}$$

Because playback levels are unknown during psychoacoustic signal analysis, the normalization procedure (Eq. 15) and the parameter PN in Eq. (16) are used to estimate SPL conservatively from the input signal. For example, a full-scale sinusoid which is precisely re-

solved by the 512-point FFT in bin k_o will yield a spectral line, $P(k_0)$, having 84 dB SPL. With 16-bit sample resolution, SPL estimates for very low amplitude input tones are lower bounded by -15 dB SPL. An example PSD estimate obtained in this manner for a CD-quality pop music selection is given in Fig. 9a. The spectrum is shown both on a linear frequency scale (upper plot) and on the bark scale (lower plot). The dashed line in both plots corresponds to the absolute threshold of hearing approximation used by the model.

STEP 2: IDENTIFICATION OF TONAL AND NOISE MASKERS

After PSD estimation and SPL normalization, tonal and non-tonal masking components are identified. Local maxima in the sample PSD which exceed neighboring components within a certain bark distance by at least 7 dB are classified as tonal. Specifically, the "tonal" set, S_T, is defined as

$$S_T = \left\{ P(k) \left| \begin{array}{l} P(k) > P(k \pm 1), \\ P(k) > P(k \pm \Delta_k) + 7\ dB \end{array} \right. \right\} \quad (18)$$

where

$$\Delta_k \in \begin{cases} 2 & 2 < k < 63 \quad (0.17\text{-}5.5 \text{ kHz}) \\ [2,3] & 63 \le k < 127 \quad (5.5\text{-}11 \text{ kHz}) \\ [2,6] & 127 \le k \le 256 \quad (11\text{-}20 \text{ kHz}) \end{cases} \quad (19)$$

Tonal maskers, $P_{TM}(k)$, are computed from the spectral peaks listed in S_T as follows

$$P_{TM}(k) = 10 \log_{10} \sum_{j=-1}^{1} 10^{0.1P(k+j)} \text{ (dB)} \quad (20)$$

Tonal maskers extracted from the example pop music selection are identified using 'x' symbols in Fig. 9a. A single noise masker for each critical band, $P_{NM}(\bar{k})$, is then computed from (remaining) spectral lines not within the $\pm\Delta_k$ neighborhood of a tonal masker using the sum

$$P_{NM}(\bar{k}) = 10 \log_{10} \sum_{j} 10^{0.1P(j)} \text{ (dB)}, \quad (21)$$

$$\forall P(j) \notin \left\{ P_{TM}(k, k \pm 1, k \pm \Delta_k) \right\}$$

where $\bar{k}$ is defined to be the geometric mean spectral line of the critical band, i.e.,

$$\bar{k} = \left(\prod_{j=l}^{u} j \right)^{1/(l-u+1)} \quad (22)$$

and l and u are the lower and upper spectral line boundaries of the critical band, respectively. Noise maskers are denoted in Fig. 9 by 'o' symbols. Dashed vertical lines are included in the bark scale plot to show the associated critical band for each masker.

STEP 3: DECIMATION AND REORGANIZATION OF MASKERS

In this step, the number of maskers is reduced using two criteria. First, any tonal or noise maskers below the absolute threshold are discarded, i.e., only maskers which satisfy

$$P_{TM,NM}(k) \ge T_q(k) \quad (23)$$

are retained, where $T_q(k)$ is the SPL of the threshold in quiet at spectral line k. In the pop music example, two high-frequency noise maskers identified during step 2 (Fig. 9a) are dropped after application of Eq. 23 (Figs. 9c-e). Next, a sliding 0.5 Bark-wide window is used to replace any pair of maskers occurring within a distance of 0.5 Bark by the stronger of the two. In the pop music example, two tonal maskers appear between 19.5 and 20.5 Barks (Fig. 9a). It can be seen that the pair is replaced by the stronger of the two during threshold calculations (Figs 9c-e). After the sliding window procedure, masker frequency bins are reorganized according to the subsampling scheme

$$P_{TM,NM}(i) = P_{TM,NM}(k) \quad (24)$$

$$P_{TM,NM}(k) = 0 \quad (25)$$

where

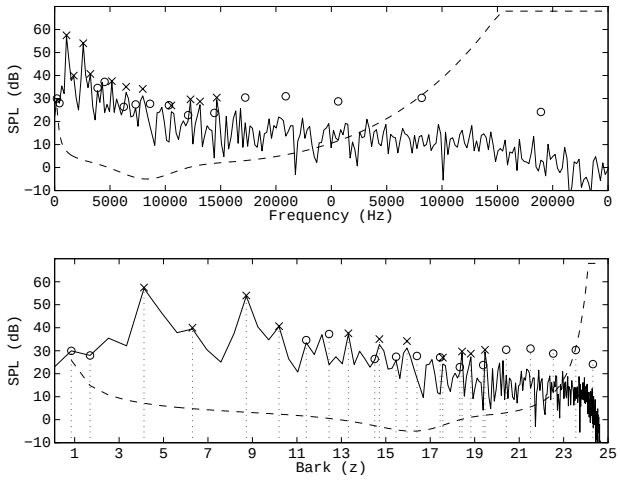
$$i = \begin{cases} k & 1 \le k \le 48 \\ k + (k \bmod 2) & 49 \le k \le 96 \\ k + 3 - ((k-1) \bmod 4) & 97 \le k \le 232 \end{cases} \quad (26)$$

The net effect of Eq. 26 is 2:1 decimation of masker bins in critical bands 18-22 and 4:1 decimation of masker bins in critical bands 22-25 , with no loss of masking components. This procedure reduces the total number of tone and noise masker frequency bins under consideration from 256 to 106. Tonal and noise maskers shown in Figs. 9c-e have been relocated according to this decimation scheme.

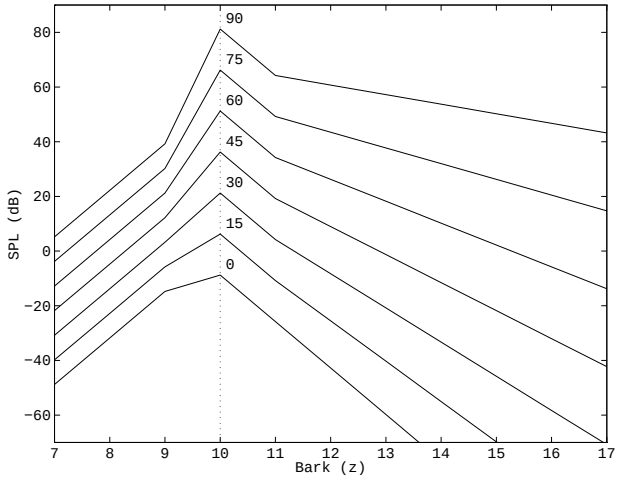
STEP 4 CALCULATION OF INDIVIDUAL MASKING THRESHOLDS

Having obtained a decimated set of tonal and noise maskers, individual tone and noise masking thresholds are computed next. Each individual threshold represents a masking contribution at frequency bin i due to the tone or noise masker located at bin j (reorganized during step 3). Tonal masker thresholds, $T_{TM}(i,j)$, are given by

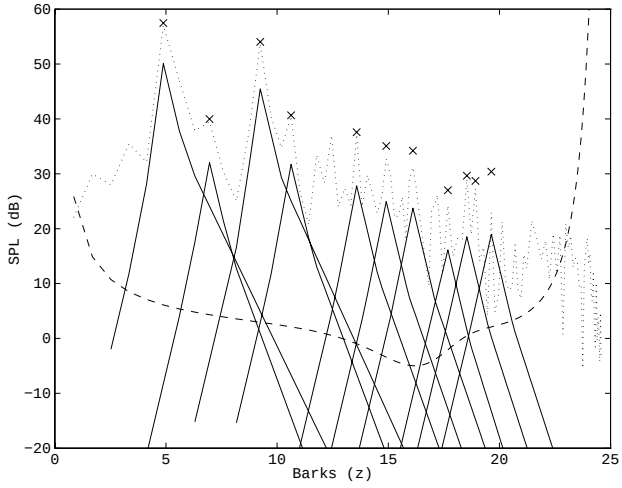
$$T_{TM}(i,j) = P_{TM}(j) - 0.275z(j) + SF(i,j) - 6.025 \text{ (dB SPL)} \quad (27)$$



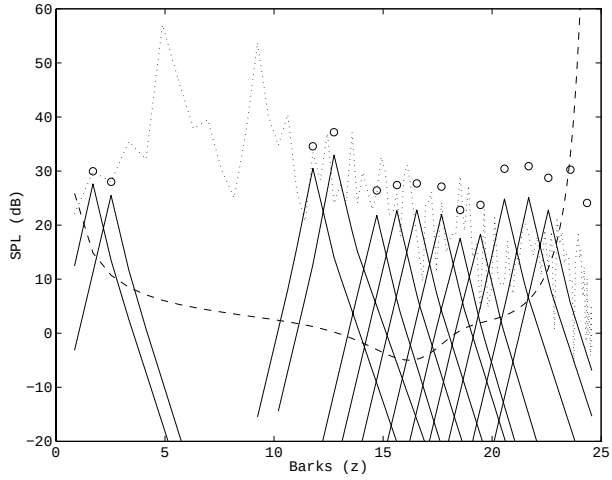
(a)



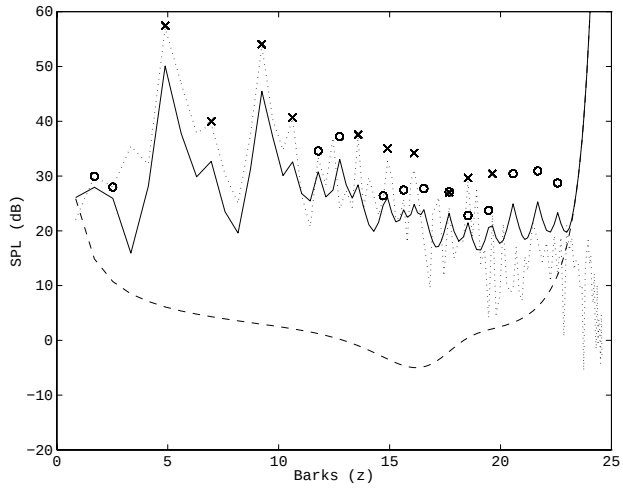
(b)



(c)



(d)



(e)

Fig. 9. Psychoacoustic Analysis for Pop Music Selection. (a) Steps 1,2: Normalized PSD, Tonal/Non-Tonal Masker ID. (b) Step 4: Prototype Spreading Functions. (c) Steps 3,4: Individual Tonal Masker Thresholds. (d) Steps 3,4: Individual Noise Masker Thresholds. (e) Step 5: Global Masking Thresholds

where $P_{TM}(j)$ denotes the SPL of the tonal masker in frequency bin j, $z(j)$ denotes the Bark frequency of bin j (Eq. 3), and the spread of masking from masker bin j to maskee bin i, $SF(i,j)$, is modeled by the expression

$$SF(i,j) = \begin{cases} 17\Delta_z - 0.4P_{TM}(j) + 11, & -3 \le \Delta_z < -1 \\ (0.4P_{TM}(j) + 6)\Delta_z, & -1 \le \Delta_z < 0 \\ -17\Delta_z, & 0 \le \Delta_z < 1 \\ (0.15P(j)_{TM} - 17)\Delta_z - 0.15P(j)_{TM}, & 1 \le \Delta_z < 8 \end{cases} \quad (28)$$

(dB SPL)

i.e., as a piecewise linear function of masker level, $P(j)$, and Bark maskee-masker separation, $\Delta_z = z(i) - z(j)$. $SF(i,j)$ approximates the basilar spreading (excitation pattern) described in section II-C. Prototype individual masking thresholds, $T_{TM}(i,j)$, are shown as a function of masker level in Fig. 9b for an example tonal masker occurring at z=10 Barks. As

shown in the figure, the slope of $T_{TM}(i,j)$ decreases with increasing masker level. This is a reflection of psychophysical test results, which have demonstrated [33] that the ear's frequency selectivity decreases as stimulus levels increase. It is also noted here that the spread of masking in this particular model is constrained to a 10-Bark neighborhood for computational efficiency. This simplifying assumption is reasonable given the very low masking levels which occur in the tails of the basilar excitation patterns modeled by $SF(i,j)$. Figure 9c shows the individual masking thresholds (Eq. 27) associated with the tonal maskers in Fig. 9a ('x'). It can be seen here that the pair of maskers identified near 19 Barks has been replaced by the stronger of the two during the decimation phase. The plot includes the absolute hearing threshold for reference. Individual noise masker thresholds, $T_{NM}(i,j)$, are given by

$$T_{NM}(i,j) = P_{NM}(j) - 0.175z(j) + SF(i,j) - 2.025 \quad \text{(dB SPL)} \tag{29}$$

where $P_{NM}(j)$ denotes the SPL of the noise masker in frequency bin j, $z(j)$ denotes the Bark frequency of bin j (Eq. 3), and $SF(i,j)$ is obtained by replacing $P_{TM}(j)$ with $P_{NM}(j)$ everywhere in Eq. 28. Figure 9d shows individual masking thresholds associated with the noise maskers identified in step 2 (Fig. 9a 'o'). It can be seen in Fig. 9d that the two high frequency noise maskers which occur below the absolute threshold have been eliminated.

STEP 5: CALCULATION OF GLOBAL MASKING THRESHOLDS

In this step, individual masking thresholds are combined to estimate a *global* masking threshold for each frequency bin in the subset given by Eq. 26. The model assumes that masking effects are additive. The global masking threshold, $T_g(i)$, is therefore obtained by computing the sum

$$T_g(i) = 10\log_{10}\left(10^{0.1T_q(i)} + \sum_{l=1}^{L} 10^{0.1T_{TM}(i,l)} + \sum_{m=1}^{M} 10^{0.1T_{NM}(i,m)}\right) \quad \text{(dB SPL)} \tag{30}$$

where $T_q(i)$ is the absolute hearing threshold for frequency bin i, $T_{TM}(i,l)$ and $T_{NM}(i,m)$ are the individual masking thresholds from step 4, and L and M are the number of tonal and noise maskers, respectively, identified during step 3. In other words, the global threshold for each frequency bin represents a signal-dependent, power additive modification of the absolute threshold due to the basilar spread of all tonal and noise maskers in the signal power spectrum. Figure 9e shows global masking threshold obtained by adding the power of the individual tonal (Fig. 9c) and noise (Fig. 9d) maskers to the absolute threshold in quiet.

III. TRANSFORM CODERS

Transform coding algorithms for high-fidelity audio make use of unitary transforms for the time/frequency analysis section in Fig. 1. These algorithms typically achieve high resolution spectral estimates at the expense of adequate temporal resolution. Many transform coding algorithms for wideband and high-fidelity audio have been proposed in the last decade. This section first describes the individual algorithms proposed by Schroeder at Thompson Consumer Electronics (MSC) [3], Brandenburg at Erlangen University (OCF) [5][43][44], Johnston at AT&T Bell Labs (PXFM/hybrid coder) [6][8], and Mahieux at Centre National d'Etudes des Telecommunications (CNET) [47][48]. Much of this work was motivated by standardization activities, and ISO/IEC eventually clustered these proposals into a single candidate algorithm, Adaptive Spectral Entropy Coding of High Quality Music Signals (ASPEC) [9], which competed successfully for inclusion in the ISO/IEC MPEG-1 [17] and MPEG-2 [18] audio coding standards. Most of MPEG-1 and MPEG-2 layer III is derived from ASPEC. Following the ASPEC discussion, the second part of this section describes novel transform coding algorithms which are not associated with ASPEC, including several very recent proposals.

The algorithms which were eventually clustered into the ASPEC proposal to ISO/IEC for MPEG audio came from researchers in both the U.S. and Europe. In Europe, some early applications of psychoacoustic principles to high fidelity audio coding were investigated by Krahe [41] during work on his dissertation [42]. Schroeder at Thompson [3] later extended these ideas into Multiple Adaptive Spectral Audio Coding (MSC). MSC utilizes a 1024-point DFT, then groups coefficients into 26 subbands which correspond to the critical bands of the ear. DFT magnitude and phase components are quantized and encoded in a two-step coarse-fine procedure which relies upon psychoacoustic bit allocation. Schroeder reported nearly transparent coding of CD-quality audio at 132 kbps [3].

A. OPTIMUM CODING IN THE FREQUENCY DOMAIN (OCF-1,OCF-2,OCF-3)

Brandenburg in 1987 proposed a 132 kbps algorithm known as Optimum Coding in the Frequency Domain (OCF) [5] which is in some respects similar to the well known Adaptive Transform Coder (ATC) for speech. OCF (Fig. 10) works as follows. The input signal is first buffered in 512 sample blocks and transformed to the frequency domain using the discrete cosine transform (DCT). Next, transform components are quan-

tized and entropy coded. A single quantizer is used for all transform components. Adaptive quantization and entropy coding work together in an iterative procedure to achieve a fixed bit rate. The initial quantizer step size is derived from the SFM (Eq. 10). In the inner loop of Fig. 10, the quantizer step size is iteratively increased and a new entropy-coded bit stream is formed at each update until the desired bit rate is achieved. Increasing the step size at each update produces fewer levels which in turn reduces the bit rate.

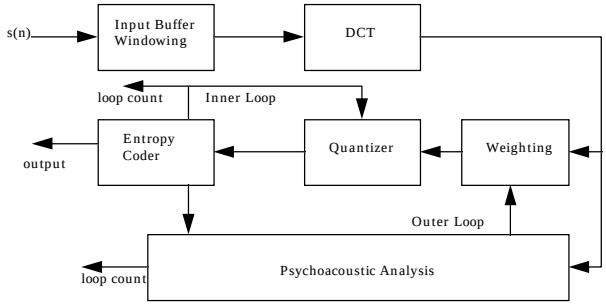


Fig. 10. OCF Encoder (after [44])

Using a second iterative procedure, psychoacoustic masking is introduced after the inner loop is done. First, critical band analysis is applied. Then, a masking function is applied which combines a flat -6 dB masking threshold with an inter-band masking threshold, leading to an estimate of JND for each critical band. If after inner loop quantization and entropy encoding the measured distortion exceeds JND in at least one critical band, quantization step sizes are adjusted in the out of tolerance critical bands *only*. The outer loop repeats until JND criteria are satisfied or a maximum loop count is reached. Entropy coded transform components are then transmitted to the receiver, along with side information which includes the log encoded SFM, the number of quantizer updates during the inner loop, and the number of step size reductions which occurred for each critical band in the outer loop. This side information is sufficient to decode the transform components and perform reconstruction at the receiver.

Brandenburg in 1988 reported an enhanced OCF (OCF-2) which achieved subjective quality improvements at a reduced bit rate of only 110 kbps [43]. The improvements were realized by replacing the DCT with the Modified DCT (MDCT - section J) and adding a pre-echo detection/compensation scheme. OCF-2 contains the first reported application of the MDCT to audio coding. Reconstruction quality is improved due to the effective time resolution increase due to the 50% time overlap associated with the MDCT. OCF-2 quality is also improved for difficult signals such as triangle and castanets due to a simple preecho detection/compensation scheme. The encoder detects pre-echos using analysis-by-synthesis. Pre-echos are detected when noise energy in a reconstructed segment (16 samples = 0.36 ms @ 44.1 kHz) exceeds signal energy. The encoder then determines the frequency below which 90% of signal energy is contained and transmits this cutoff to the decoder. Given pre-echo detection at the encoder (1 bit) and a cutoff frequency, the decoder discards frequency components above the cutoff, in effect lowpass filtering preechoes. Due to these enhancements, OCF-2 was reported to achieve transparency over a wide variety of source material. Only some experts were able to detect pre-echo distortion in difficult signals such as the glockenspiel. Later in 1988 Brandenburg reported further OCF enhancements (OCF-3) in which he reported better quality at a lower bit rate (64 kbps) with reduced complexity [44]. OCF-3 benefited from several improvements relative to OCF-2. First, differential coding was applied to spectral components to exploit correlation between adjacent samples. Second, the psychoacoustic model was modified to account for temporal pre- and post-masking. Third, errors in the OCF-2 quantizer were identified and corrected. Finally, step size coarseness for the inner quantization loop was increased in OCF-3, resulting in reduced complexity.

B. PERCEPTUAL TRANSFORM CODER (PXFM)

While Brandenburg developed OCF, similar work was simultaneously underway at AT&T Bell Labs. James Johnston [6] developed several DFT-based transform coders for audio during the late eighties which became an integral part of the ASPEC proposal. Johnston's work in perceptual entropy forms the basis for a 4(3)-bit/sample transform coder reported in 1988 [6] which achieves transparent coding of FM-quality monaural audio signals (Fig. 11). The idea behind the perceptual transform coder (PXFM) is to estimate the amount of quantization noise which can be inaudibly injected into each transform domain subband using PE estimates. The coder is memoryless and works as follows. The signal is first windowed into overlapping (1/16) segments and transformed using a 2048-point FFT. Next, the PE procedure described in section one is used to estimate JND thresholds for each critical band. Then, an iterative quantization loop adapts a set of 128 subband quantizers to satisfy the JND thresholds until the fixed bit rate is achieved. Finally, quantization and bit packing are performed. Quantized transform components are transmitted to the receiver along with appropriate side information. Quantization subbands consist of 8-sample blocks of complex-valued transform components.

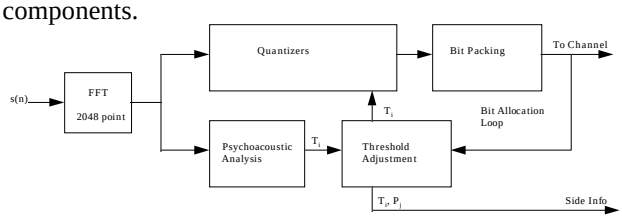


Fig. 11. PXFM Encoder (after [6])

The quantizer adaptation loop first initializes the $j \in [1,128]$ subband quantizers (1024 unique FFT components/8 components per subband) with k_j levels and step sizes of T_i as follows:

$$k_j = 2 * \text{nint}\left(\frac{P_j}{T_i}\right) + 1 \qquad (31)$$

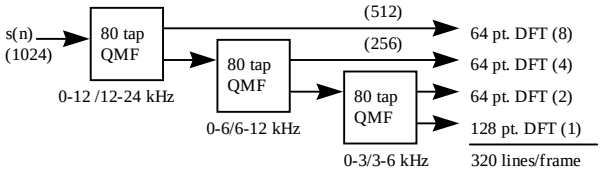
where T_i are the quantized critical band JND thresholds, P_j is the quantized magnitude of the largest real or imaginary transform component in the *jth* subband, and *nint*() is the nearest integer rounding function. The adaptation process involves repeated application of two steps. First, bit packing is attempted using the current quantizer set. Although many bit packing techniques are possible, one simple scenario involves sorting quantizers in k_j order, then filling 64-bit words with encoded transform components according to the sorted results. After bit packing, T_i are adjusted by a carefully controlled scale factor, and the adaptation cycle repeats. Quantizer adaptation halts as soon as the packed data length satisfies the desired bit rate. Both P_j and the modified T_i are quantized on a dB scale using 8-bit uniform quantizers with a 170 dB dynamic range. These parameters are transmitted as side information and used at the receiver to recover quantization levels (and thus implicit bit allocations) for each subband, which are in turn used to decode quantized transform components. The DC FFT component is quantized with 16 bits and is also transmitted as side information.

In 1989, Johnston extended the PXFM coder to handle stereophonic signals (SEPXFM) and attained transparent coding of a CD-quality stereophonic channel at 192 kb/s, or 2.2 bits/sample. SEPXFM [45] realizes performance improvements over PXFM by exploiting inherent stereo cross-channel redundancy and by assuming that both channels are presented to a single listener rather than being used as separate signal sources. SEPXFM structure is similar to that of PXFM, with variable radix bit packing replaced by adaptive entropy coding. Side information is therefore reduced to include only adjusted JND thresholds (step-sizes) and pointers to the entropy codebooks used in each transform domain subband. The coder works in the following manner. First, sum (*L*+*R*) and difference (*L*-*R*) signals are extracted from the left (*L*) and right (*R*) channels to exploit left/right redundancy. Next, the sum and difference signals are windowed and transformed using the FFT. Then, a single JND threshold for each critical band is established via the PE method using the summed power spectra from the *L*+*R* and *L*-*R* signals. A single combined JND threshold is applied to quantization noise shaping for both signals (*L*+*R* and *L*-*R*), based upon the assumption that a listener is more than one "critical distance" [46] from away from the stereo speakers. Like PXFM, a fixed bit rate is achieved by applying an iterative threshold adjustment procedure after the initial determination of JND levels. The adaptation process, analogous to PXFM bit rate adjustment and bit packing, consists of several steps. First, transform components from both *L*+*R* and *L*-*R* are split into subband blocks, each averaging 8 real/imaginary samples. Then, one of six entropy codebooks is selected for each subband based on the average component magnitude within that subband. Next, transform components are quantized given the JND levels and encoded using the selected codebook. Subband codebook selections are themselves entropy encoded and transmitted as side information. After encoding, JND thresholds are scaled by an estimator and the quantizer adaptation process repeats. Threshold adaptation stops when the combined bitstream of quantized JND levels, Huffman encoded *L*+*R* components, Huffman encoded *L*-*R* components, and Huffman encoded average magnitudes achieves the desired bit rate. The Huffman codebooks are developed using a large music and speech database. They are optimized for difficult signals at the expense of mean compression rate. It is also interesting to note that headphone listeners reported no noticeable acoustic mixing, despite the critical distance assumption and single combined JND level estimate for both channels, *L*+*R* and *L*-*R*.

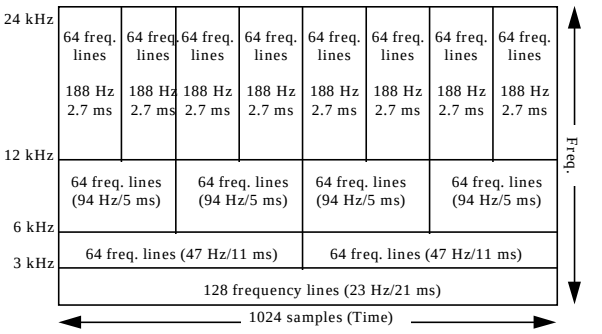
C. AT&T HYBRID CODER

Following the success of their individual coders, Johnston and Brandenburg [8] collaborated in 1990 to produce a hybrid coder which, strictly speaking, is both a subband and transform algorithm. It is included in this section because it was part of the ASPEC cluster. The idea behind the hybrid coder is to improve time and frequency resolution relative to OCF and PXFM by constructing a filterbank which more closely resembled the human ear. This is accomplished at the encoder by first splitting the input signal into four octave-width subbands using a QMF filterbank. The decimated output sequence from each subband is then followed by one or more transforms to achieve the desired time/frequency resolution (Fig. 12a). Both DFT and MDCT transforms were investigated. Given the tiling of the time-frequency plane shown in Fig. 12b, frequency resolution at low frequencies (23.4 Hz) is well matched to the ear, while the time resolution at high frequencies (2.7 ms) is sufficient for pre-echo control. The quantization and coding schemes of the hybrid coder combine elements from both PXFM and OCF. Masking thresholds are estimated using the PXFM approach for eight time slices in each frequency subband. A more sophisticated tonality estimate was defined to replace the SFM (Eq. 10) used in PXFM, however, such that tonality is estimated in the hybrid coder as a local characteristic of each individual spectral line. Predict-

ability of magnitude and phase spectral components across time is used to evaluate tonality instead of just global spectral shape within a single frame. High temporal predictability of magnitudes and phases is associated with the presence of a tonal signal and



(a)



(b)

Fig. 12. Johnston/Brandenburg Hybrid Coder. (a) Filterbank Structure, (b) Time/Freq Tiling (after [8])

visa-versa. The hybrid coder employs a quantization and coding scheme borrowed from OCF. As far as quality, the hybrid coder without any explicit pre-echo control mechanism was reported to achieve quality better than or equal to OCF-3 at 64 kbps [8]. The only disadvantage noted by the authors was increased complexity. A similar hybrid structure was eventually adopted in MPEG-1 and -2 Layer III.

D. CNET CODER

During the same period in which Schroeder, Brandenburg, and Johnston pursued optimal transform coding algorithms for audio, so to did researchers at CNET. In 1989, Mahieux, Petit, *et al.* proposed a DFT-based audio coding system which introduced a novel scheme to exploit DFT interblock redundancy. They reported nearly transparent quality for 15 kHz (FM-grade) audio at 96 kbps [47], except for some highly harmonic signals. The encoder applies first-order backward-adaptive predictors (across time) to DFT magnitude and differential phase components, then quantizes separately the prediction residuals. Magnitude and differential phase residuals are quantized using an adaptive non-uniform pdf-optimized quantizer designed for a Laplacian distribution and an adaptive uniform quantizer, respectively. The backward-adaptive quantizers are reinitialized during transients. Bits are allocated during step-size adaptation to shape quantization noise such that a psychoacoustic noise threshold is satisfied for each block. The psychoacoustic model used is similar to Johnston's model described in section II. The use of linear prediction is justified because it exploits magnitude and differential phase time redundancy, which tends to be large during periods when the audio signal is quasi-stationary, especially for signal harmonics. Quasi-stationarity might occur, for example, during a sustained note.

In 1990, Mahieux and Petit reported on the development of an MDCT-based transform coder for which they claimed transparent CD-quality at 64 kbps [48]. This algorithm introduced a novel "spectrum descriptor" scheme for representing the power spectral envelope. The algorithm first segments input audio into frames of 1024 samples, corresponding to 12 msec of new data per frame, given 50% MDCT time overlap. Then, bit allocation is computed at the encoder using a set of "spectrum descriptors." Spectrum descriptors consist of quantized sample variances for MDCT coefficients grouped into 35 non-uniform frequency subbands. Like their DFT coder, this algorithm exploits either interblock or intrablock redundancy by differentially encoding the spectrum desciptors with respect to time or frequency and transmitting them to the receiver as side information. A decision whether to code with respect to time or frequency is made on the basis of which method requires fewer bits; the binary decision requires only 1 bit. Either way, spectral descriptor encoding is done using log DPCM with a first-order predictor and a 16-level uniform quantizer with a step-size of 5 dB. Huffman coding of the spectral descriptor codewords results in less than 2-bits/descriptor. A global masking threshold is estimated by convolving the spectral descriptors with a basilar spreading function on a bark scale, somewhat like the approach taken by Johnston' s PXFM. Bit allocations for quantization of normalized transform coefficients are obtained from the masking threshold estimate. As usual, bits are allocated such that quantization noise is below the masking threshold at every spectral line. Transform coefficients are normalized by the appropriate spectral descriptor, then quantized and coded, with one exception. Masked transform coefficients, which have lower energy than the global masking threshold, are treated differently. The authors found that masked coefficient bins tend to be clustered, therefore they can be compactly represented using run length encoding (RLE). RLE codewords are Huffman coded for maximum coding gain. The coder was reported to perform well for broadband signals with many harmonics but had some problems in the case of spectrally flat signals.

More recently, Mahieux and Petit enhanced their 64 kbps algorithm by incorporating a sophisticated pre-echo detection and postfiltering scheme, as well as incorporating a novel quantization scheme for 2-coefficient (low-frequency) spectral descriptor bands [104]. For improved quantization performance, two-component spectral descriptors are efficiently vector

encoded in terms of polar coordinates. Pre-echos are detected at the encoder and flagged using 1 bit. The idea behind the pre-echo compensation is to temporarily activate a postfilter at the decoder in the corrupted quiet region prior to the signal attack, therefore a stopping index must also be transmitted. The 2nd-order IIR postfilter difference equation is given by

$$\hat{s}_{pf}(n) = b_0\hat{s}(n) + a_1\hat{s}_{pf}(n-1) + a_2\hat{s}_{pf}(n-2) \qquad (32)$$

where $\hat{s}(n)$ is the non-postfiltered output signal which is corrupted by pre-echo distortion, $\hat{s}_{pf}(n)$ is the postfiltered output signal, and a_i are related to the parameters α_i by

$$a_1 = \alpha_1\left[1-\left(\frac{p(0,0)}{p(0,0)+\sigma_b^2}\right)\right], \qquad (33a)$$

$$a_2 = \alpha_2\left[1-\left(\frac{p(1,0)}{p(0,0)+\sigma_b^2}\right)\right] \qquad (33b)$$

where α_i are the parameters of a 2nd-order autoregressive (AR-2) spectral estimate of the output audio, $\hat{s}(n)$, during the previous non-postfiltered frame. The AR-2 estimate, $\dot{s}(n)$, can be expressed in the time domain as

$$\dot{s}(n) = w(n) + \alpha_1\dot{s}(n-1) + \alpha_2\dot{s}(n-2) \qquad (34)$$

where $w(n)$ represents gaussian white noise. The prediction error is then defined as

$$e(n) = \hat{s}(n) - \dot{s}(n) \qquad (35)$$

The parameters $p(i,j)$ in Eq. 33a and 33b are elements of the prediction error covariance matrix, $\mathbf{P}$, and the parameter σ_b^2 is the pre-echo distortion variance, which is derived from side information. Pre-echo postfiltering and improved quantization schemes resulted in a subjective score of 3.65 for two-channel stereo coding at 64 kbps per channel on the 5-point CCIR 5-grade impairment scale over a wide range of listening material. The CCIR J.41 reference audio codec (MPEG-1, Layer-II) achieved a score of 3.84 at 384 kbps/channel over the same set of tests.

E. ASPEC

The MSC, OCF, PXFM, AT&T hybrid, and CNET audio transform coders were eventually clustered into a single proposal by the ISO/IEC JTC1/SC2 WG11 committee. As a result, Schroeder, Brandenburg, Johnston, Herre, and Mahieux collaborated in 1991 to propose for acceptance as the new MPEG audio compression standard a flexible coding algorithm, ASPEC, which incorporated the best features of each coder in the group [9]. ASPEC was claimed to produce better quality than any of the individual coders at 64 kbps. The structure of ASPEC combines elements from all of its predecessors. Like OCF and the CNET coder, ASPEC uses the MDCT for time-frequency mapping. The masking model is similar to that used in PXFM and the AT&T hybrid coder, including the sophisticated tonality estimation scheme at lower bit rates. The quantization and coding procedures use the pair of nested loops proposed for OCF, as well as the block differential coding scheme developed at CNET. Moreover, long runs of masked coefficients are run-length and Huffman encoded. Quantized scalefactors and transform coefficients are Huffman coded also. Pre-echos are controlled using a dynamic window switching mechanism, like the Thompson coder. ASPEC offers several modes for different quality levels, ranging from 64 to 192 kbps per channel. A real-time ASPEC implementation for coding one channel at 64 kbps was realized on a pair of 33 MHz Motorola DSP56001 devices. ASPEC ultimately formed the basis for Layer III of the MPEG-1 and MPEG-2 standards. We note here that similar contributions have been made in the area of transform coding for audio outside the ASPEC cluster. For example, Iwadare, *et al.* reported on DCT-based [49] and MDCT-based [11] perceptual adaptive transform coders which control pre-echo distortion using adaptive window size.

F. DPAC

Other investigators have also developed promising schemes for transform coding of audio. Paraskevas and Mourjopoulos [106] reported on a differential perceptual audio coder (DPAC) which makes use of a novel scheme for exploiting long-term correlations. DPAC works as follows. Input audio is transformed using the MDCT. A two-state classifier then labels each new frame of transform coefficients as either a "reference" frame or a "simple" frame. The classifier labels as "reference" frames which contain significant audible differences from the previous frame. The classifier labels non-reference frames as "simple." Reference frames are quantized and encoded using scalar quantization and psychoacoustic bit allocation strategies similar to Johnston's PXFM. Simple frames, however, are subjected to coefficient substitution. Coefficients whose magnitude differences with respect to the previous reference frame are below an experimentally optimized threshold are replaced at the decoder by the corresponding reference frame coefficients. The encoder, then, replaces subthreshold coefficients with zeros, thus saving transmission bits. Unlike the interframe predictive coding schemes of Mahieux and Petit, the DPAC coefficient substitution system is advantageous in that it guarantees the "simple" frame bit allocation will always be less than or equal to the bit allocation which would be required if the frame was coded as a "reference" frame. Superthreshold "simple" frame coefficients are coded in the same way as reference frame coefficients. DPAC performance was evaluated for frame classifiers which utilized three different selection criterion. Under the Euclidean criterion, test frames satisfying the inequality

$$\left[\frac{\mathbf{s}_d{}^T\mathbf{s}_d}{\mathbf{s}_r{}^T\mathbf{s}_r}\right]^{\frac{1}{2}} \le \lambda \qquad (36)$$

are classified as "simple", where the vectors $\mathbf{s}_r$ and $\mathbf{s}_t$, respectively, contain reference and test frame time-domain samples, and the difference vector, $\mathbf{s}_d$, is defined as

$$\mathbf{s}_d = \mathbf{s}_r - \mathbf{s}_t \qquad (37)$$

Under the PE (Eq. 14) criterion, a test frame is labeled as "simple" if it satisfies the inequality

$$\frac{PE_S}{PE_R} \le \lambda \qquad (38)$$

where PE_S corresponds to the PE of the "simple" (coefficient-substituted) version of the test frame, and PE_R corresponds to the PE of the unmodified test frame. Finally, under the SFM (Eq. 10) criterion, a test frame is labeled as "simple" if it satisfies the inequality

$$abs\left(10\log_{10}\frac{SFM_T}{SFM_R}\right) \le \lambda \qquad (39)$$

where SFM_T corresponds to the test frame SFM, and SFM_R corresponds to the SFM of the previous reference frame. The decision threshold, λ, was experimentally optimized for all three criteria. Best performance was obtained while encoding source material using the PE criterion. As far as overall performance is concerned, noise-to-mask ratio (NMR) measurements were compared between DPAC and Johnston's PXFM algorithm at 64, 88, and 128 kbps. Despite an average drop of 30-35% in PE measured at the DPAC coefficient substitution stage output relative to the coefficient substitution input, comparative NMR studies indicated that DPAC outperforms PXFM only below 88 kbps and then only for certain types of source material such as pop or jazz music. The desirable PE reduction led to an undesirable drop in reconstruction quality. The authors concluded that DPAC may be preferable to algorithms such as PXFM for low bit rate, non-transparent applications.

G. DFT NOISE SUBSTITUTION

Other coefficient substitution schemes have also been proposed. Whereas DPAC exploits temporal correlation, a substitution technique which exploits decorrelation was recently devised for coding efficiently noise-like portions of the spectrum. In a noise substitution procedure [50], Schulz parameterizes transform coefficients corresponding to noise-like portions of the spectrum in terms of average power, frequency range, and temporal evolution, resulting in an increased coding efficiency of 15% on average. A temporal envelope for each parametric noise band is required because transform block sizes for most codecs are much longer (e.g., 30 ms) than the human auditory system's temporal resolution (e.g., 2 ms). In this method, noise-like spectral regions are identified in the following way. First, least-mean-square (LMS) adaptive linear predictors (LP) are applied to the output channels of a multi-band QMF analysis filterbank which has as input the original audio, $s(n)$. A predicted signal, $\hat{s}(n)$, is obtained by passing the LP output sequences through the QMF synthesis filterbank. Prediction is done in subbands rather than over the entire spectrum to prevent classification errors which could result if high-energy noise subbands are allowed to dominate predictor adaptation, resulting in misinterpretation of low-energy tonal subbands as noisy. Next, the DFT is used to obtain magnitude ($S(k), \hat{S}(k)$) and phase components ($\theta(k), \hat{\theta}(k)$), of the input, $s(n)$, and prediction, $\hat{s}(n)$, respectively. Then, tonality, $T(k)$, is estimated as a function of the magnitude and phase predictability, i.e,

$$T(k) = \alpha\left|\frac{S(k)-\hat{S}(k)}{S(k)}\right| + \beta\left|\frac{\theta(k)-\hat{\theta}(k)}{\theta(k)}\right| \qquad (40)$$

where α and β are experimentally determined constants. Noise substitution is applied to contiguous blocks of transform coefficient bins for which $T(k)$ is very small. The 15% average bit savings realized using this method in conjunction with transform coding is offset to a large extent by a significant complexity increase due to the additions of the adaptive linear predictors and a multi-band analysis-synthesis QMF filterbank. As a result, the author and focused his attention on the application of noise substitution to QMF-based subband coding algorithms.

H. DCT WITH VECTOR QUANTIZATION

For the most part, the algorithms described thus far rely upon scalar quantization of transform coefficients. This is not unreasonable, since scalar quantization in combination with entropy coding can achieve very good performance. As one might expect, however, vector quantization (VQ) has also been applied to transform coding of audio, although on a much more limited scale. For example, Gersho and Chan investigated several VQ schemes for coding DCT coefficients subject to a constraint of minimum perceptual distortion. They first reported on a variable rate coder [7] which achieves high quality in the range of 55 to 106 kbps for audio sequences bandlimited to 15 kHz (32 kHz sample rate). After computing the DCT on 512 sample blocks, the algorithm utilizes a novel Multi-Stage Tree-Structured VQ (MSTVQ) scheme for quantization of normalized vectors, with each vector containing 4 DCT components. Bit allocation and vector normalization are derived at both the encoder and decoder from a sampled power spectral envelope which consists of 29 groups of transform coefficients. A simplified masking model assumes that each sample of the power envelope repre-

sents a single masker. Masking is assumed to be additive, as in the ASPEC algorithms. Thresholds are computed as a fixed offset from the masking level. The authors observed a strong correlation between SFM and the amount of offset required to achieve high quality. Two-segment scalar quantizers that are piecewise linear on a dB scale are used to encode the power spectral envelope. Quadratic interpolation is used to restore full resolution to the subsampled envelope.

Gersho and Chan later enhanced [51] their algorithm by improving the power envelope and transform coefficient quantization schemes. In the new approach to quantization of transform coefficients, constrained-storage VQ [52] techniques (CS-VQ) are combined with the MSTVQ (CS-MSTVQ) from the original coder, allowing the new coder to handle peak Noise-to-Mask ratio (NMR) requirements without impractical codebook storage requirements. In fact, CS-MSTVQ enabled quantization of 127 4-coefficient vectors using only 4 unique quantizers. Power spectral envelope quantization is enhanced by extending its resolution is extended to 127 samples. The samples are then encoded using a two-stage process. The first stage applies nonlinear interpolative VQ (NLIVQ), a dimensionality reduction process which represents the 127-element power spectral envelope vector using only a 12-dimensional "feature power envelope." Unstructured VQ is applied to the feature power envelope. Then, a full-resolution quantized envelope is obtained from the unstructured VQ index into a corresponding interpolation codebook. In the second stage, segments of the envelope residual are encoded using a set of 8-, 9-, and 10-element TSVQ quantizers. Relative to their first VQ/DCT coder, the authors reported savings of 10-20 kbps with no reduction in quality due to the CS-VQ and NLIVQ schemes.

I. MDCT WITH VECTOR QUANTIZATION

More recently, Iwakami *et al.* developed Transform-Domain Weighted Interleave Vector Quantization (TWIN-VQ), an MDCT-based coder which also involves transform coefficient VQ [105]. This algorithm exploits LPC analysis, spectral inter-frame redundancy, and interleaved VQ. At the encoder (Fig. 13), each frame of MDCT coefficients is first divided by the corresponding elements of the LPC spectral envelope, resulting in a spectrally flattened quotient (residual) sequence. This procedure flattens the MDCT envelope but does not affect the fine structure. The next step, therefore, divides the first step residual by a predicted fine structure envelope. This predicted fine structure envelope is computed as a weighted sum of three previous quantized fine structure envelopes, i.e., using backward prediction. Interleave VQ is applied to the normalized second step residual. The interleave VQ vectors are structured in the following way. Each N-sample normalized second step residual vector is split into K subvectors, each containing N/K coefficients. Second step residuals from the N-sample vector are interleaved in the K subvectors such that the i^{-th} subvector contains elements i+nK, where n=0,1,...,(N/K)-1. Perceptual weighting is also incorporated by weighting each subvector by a non-linearly transformed version of its corresponding LPC envelope component prior to the codebook search. VQ indices are transmitted to the receiver. Side information consists of VQ normalization coefficients and the LPC envelope encoded in terms of LSPs. The authors claim higher subjective quality than MPEG-1 Layer II at 64 kbps for 48 kHz CD-quality audio, as well as higher quality than MPEG-1 Layer II for 32 kHz audio at 32 kbps. Enhancements to the weighted interleaving scheme and LPC envelope representation are reported in [53] which enabled real-time implementation of stereo decoders on Pentium and PowerPC platforms. Channel error robustness issues are addressed in [54].

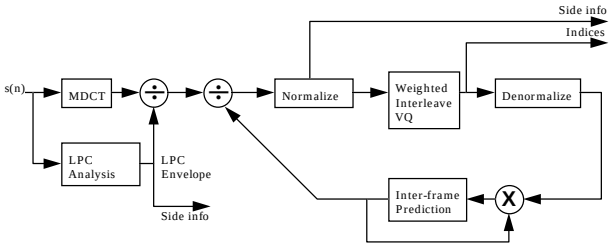


Fig. 13. TWIN-VQ Encoder (after [105])

J. MODIFIED DISCRETE COSINE TRANSFORM (MDCT)

Before concluding the transform coder discussion and embarking upon consideration of subband algorithms, it is useful to consider briefly the modified discrete cosine transform (MDCT), a recently developed modulated lapped transform which has found widespread application throughout the audio coding literature. Several of the algorithms discussed in sections III, IV, and V make use of this transform. The MDCT offers the advantage of overlapping time windows while managing to preserve critical sampling. The analysis window must be carefully designed such that the time-domain aliasing introduced by 50% overlap and 2:1 decimation will cancel in the inverse transformation [55]. The MDCT analysis expression is

$$X(k) = \sum_{n=0}^{2N-1} h(n)x(n)\cos\left[\frac{\pi}{2N}(2k+1)(2n+1+N)\right], \quad k = 0,1,\ldots,2N-1 \tag{41}$$

where the analysis window must satisfy

$$h^2(N-1-n)+h^2(n)=2, \quad 0 \le n < N \tag{42a}$$

$$h^2(N+n)+h^2(2N-1-n) = 2, \quad 0 \le n < N \tag{43b}$$

An example analysis window which produces the desired time-domain aliasing cancellation is given by

$$h(n) = \pm\sqrt{2}\sin\left[\left(n+\frac{1}{2}\right)\frac{\pi}{2N}\right] \qquad (44)$$

The development of FFT-based fast algorithms for the MDCT (e.g.,[56]) has made it viable for real-time applications. Coders such as ISO/MPEG, Dolby's AC-3, and Sony's ATRAC for MiniDisc make use of the MDCT.

IV. SUBBAND CODERS

Like the transform coders described in the previous section, subband coders also exploit signal redundancy and psychoacoustic irrelevancy in the frequency domain. Instead of unitary transforms, however, these coders rely upon frequency-domain representations of the signal obtained from banks of bandpass filters. The audible frequency spectrum (20 Hz - 20 kHz) is divided into frequency subbands using a bank of bandpass filters. The output of each filter is then sampled and encoded. At the receiver, the signals are demultiplexed, decoded, demodulated, and then summed to reconstruct the signal. Audio subband coders realize coding gains by efficiently quantizing and encoding the decimated output sequences from perfect reconstruction filterbanks. Efficient quantization methods usually rely upon psychoacoustically controlled dynamic bit allocation rules which allocate bits to subbands in such a way that the reconstructed output signal is free of audible quantization noise or other artifacts. In a generic subband audio coder, the input signal is first split into several uniform or non-uniform subbands using some critically sampled, perfect reconstruction filterbank. Non-ideal reconstruction properties in the presence of quantization noise are compensated for by utilizing subband filters which have very good sidelobe attenuation, an approach which usually requires high-order filters. Then, decimated output sequences from the filterbank are normalized and quantized over short, 2-to-10 millisecond (ms) blocks. Psychoacoustic signal analysis is used to allocate an appropriate number of bits for the quantization of each subband. The usual approach is to allocate a just-sufficient number of bits to mask quantization noise in each block while simultaneously satisfying some bit rate constraint. Since masking thresholds and hence bit allocation requirements are time-varying, buffering is often introduced to match the coder output to a fixed rate. The encoder sends to the decoder quantized subband output samples, normalization scalefactors for each block of samples, and bit allocation side information. Bit allocation may be transmitted as explicit side information, or it may be implicitly represented by some parameter such as the scalefactor magnitudes. The decoder uses side information and scalefactors in conjunction with an inverse filterbank to reconstruct a coded version of the original input.

Numerous subband coding algorithms for hi fidelity audio have appeared in the literature since the late eighties. This section focuses upon the individual subband algorithms proposed by researchers from the Institut fur Rundfunktechnik (IRT) [4][60], Philips Research Laboratories [61], and CCETT. Much of this work was motivated by standardization activities for the European Eureka-147 digital broadcast audio (DBA) system. The ISO/IEC eventually clustered the IRT, Philips, and CCETT proposals into a single candidate algorithm, Masking Pattern Adapted Universal Subband Integrated Coding and Multiplexing (MUSICAM) [10][62], which competed successfully for inclusion in the ISO/IEC MPEG-1 and MPEG-2 audio coding standards. Consequently, most of MPEG-1 [17] and MPEG-2 [18] layers I and II are derived from MUSICAM. Other subband algorithms were also proposed by Charbonnier and Petit [57], Voros [58], and Teh *et al.* [59], are not discussed here. The section concentrates upon MUSICAM and its antecedents, which ultimately led to the creation of the MPEG audio standard.

A. MASCAM

The MUSICAM algorithm is derived from coders developed at IRT, Philips, and CNET. At IRT, Theile, Stoll, and Link developed Masking Pattern Adapted Subband Coding (MASCAM), a subband audio coder [4] based upon a tree-structured quadrature mirror filter (QMF) filterbank which was designed to mimic the critical band structure of the auditory filterbank. The coder has 24 non-uniform subbands, with bandwidths of 125 Hz below 1 kHz, 250 Hz in the range 1-2 kHz, 500 Hz in the range 2-4 kHz, 1 kHz in the range 4-8 kHz, and 2 kHz from 8 kHz to 16 kHz. The prototype QMF filter has 64 taps. Subband output sequences are processed in 2-ms blocks. A normalization scalefactor is quantized and transmitted for each block from each subband. Subband bit allocations are derived from a simplified psychoacoustic analysis. The original coder reported in [4] considered only in-band simultaneous masking. Later, as described in [60], inter-band simultaneous masking and temporal masking were added to the bit rate calculation. Temporal postmasking is exploited by updating scalefactors less frequently during periods of signal decay. The MASCAM coder was reported to achieve high-quality results for 15 kHz bandwidth input signals at bit rates between 80 and 100 kbps per channel. A similar subband coder was developed at Philips during this same period. As described by Velhuis *et al.* in [61], the Philips group investigated subband schemes based on 20- and 26-band non-uniform filterbanks. Like the original MASCAM system, the Philips coder relies upon a highly simplified masking model which considers only the upward spread of simultaneous masking. Thresholds are derived from a prototypical basilar excitation function under worst-case assumptions regarding the frequency separation of masker and maskee. Within each subband, signal en-

ergy levels are treated as single maskers. Given SNR targets due to the masking model, uniform ADPCM is applied to the normalized output of each subband. The Philips coder was claimed to deliver high quality coding of CD-quality signals at 110 kbps for the 26-band version and 180 kbps for the 20-band version.

B. MUSICAM

Based primarily upon coders developed at IRT and Phillips, the MUSICAM algorithm [10][62] was successful in the ISO/IEC competition [63] for a new audio coding standard. It eventually formed the basis for MPEG-1 and MPEG-2 audio layers I and II. Relative to its predecessors, MUSICAM (Fig. 14) makes several practical tradeoffs between complexity, delay, and quality. By utilizing a uniform bandwidth, 32-band polyphase filterbank instead of a tree structured QMF filterbank, both complexity and delay are greatly reduced relative to the IRT and Phillips coders. Delay and complexity are 10.66 ms and 5 MFLOPS, respectively. These improvements are realized at the expense of using a sub-optimal filterbank, however, in the sense that filter bandwidths (constant 750 Hz for 48 kHz sample rate) no longer correspond to the critical band rate. Despite these excessive filter bandwidths at low frequencies, high quality coding is still possible with MUSICAM due to its enhanced psychoacoustic analysis. High resolution spectral estimates (46 Hz/line at 48 kHz sample rate) are obtained through the use of a 1024-point FFT in parallel with the polyphase filterbank. This parallel structure allows for improved estimation of masking thresholds and hence determination of more accurate minimum signal-to-mask ratios (SMRs) required within each subband. The MUSICAM psychoacoustic analysis procedure is essentially the same as the MPEG-1 psychoacoustic model 1 described in section II-G. The remainder of MUSICAM works as follows. Subband output sequences are processed in 8 ms blocks (12 samples at 48 kHz), which is close to the temporal resolution of the auditory system (4-6 ms). Scalefactors are extracted from each block and encoded using 6-bits over a 120 dB dynamic range. Occasionally, temporal redundancy is exploited by repetition over 2 or 3 blocks (16 or 24 ms) of slowly-changing scalefactors within a single subband. Repetition is avoided during transient periods such as sharp attacks. Subband samples are quantized and coded in accordance with SMR requirements for each subband as determined by the psychoacoustic analysis. Bit allocations for each subband are transmitted as side information. On the CCIR five-grade impairment scale, MUSICAM scored 4.6 (std dev. 0.7) at 128 kbps, and 4.3 (std dev. 1.1) at 96 kbps per monaural channel, compared to 4.7 (std dev. 0.6) on the same scale for the uncoded original. Quality was reported to suffer somewhat at 96 kbps for critical signals which contained sharp attacks (e.g., triangle, castanets), and this was reflected in a relatively high standard deviation of 1.1. MUSICAM was selected by ISO/IEC for MPEG audio due to its desirable combination of high quality, reasonable complexity, and manageable delay. Also, bit error robustness was found to be very good (errors nearly imperceptible) up to a bit error rate of 10^{-3}.

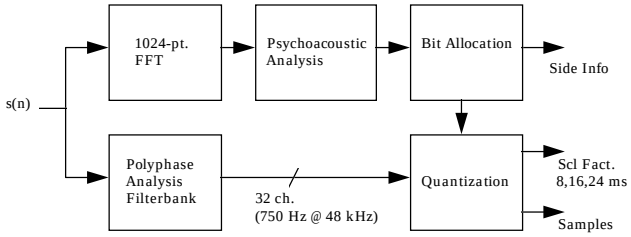


Fig. 14. MUSICAM Encoder (after [62])

C. WAVELET DECOMPOSITIONS

The previous section described subband coding algorithms which utilize banks of fixed resolution bandpass QMF or polyphase finite impulse response (FIR) filters. This section describes a different class of subband coders which rely instead upon a filterbank interpretation of the discrete wavelet transform (DWT). DWT based subband coders offer increased flexibility over the subband coders described previously since identical filterbank magnitude frequency responses can be obtained for many different choices of a wavelet basis. This flexibility presents an opportunity for basis optimization. In the context of audio coding, a desired filterbank magnitude response can first be established. This response might be matched to the auditory filterbank, for example. Then, for each segment of audio, one can adaptively choose a wavelet basis which minimizes the number of bits required to encode the signal subbands at some target distortion level. Given a psychoacoustically derived distortion target, the encoding remains perceptually transparent.

A detailed discussion of specific technical conditions associated with the various wavelet families is beyond the scope of this paper, and this section therefore avoids mathematical development and concentrates instead upon high-level coder architectures. In-depth technical information regarding wavelets is available in many references, for example [64]. Before describing the wavelet based coders, however, it is useful to summarize some basic wavelet characteristics. Wavelets are a family of basis functions for the space of square integrable signals. A finite energy signal can be represented as a weighted sum of the translates and dilates of a single wavelet. Continuous-time wavelet signal analysis can be extended to discrete time and square summable sequences. Under certain assumptions, the DWT acts as an orthonormal linear transform $T: R^N \rightarrow R^N$. For a compact (finite) support wavelet of length K, the associated transformation matrix, $\mathbf{Q}$, is

fully determined by a set of coefficients $\{c_k\}$ for $0 \le k \le K-1$. As shown in Fig. 15, this transformation matrix has an associated filterbank interpretation. One application of the transform matrix, $\mathbf{Q}$, to an $N\,x\,1$ signal vector, $\mathbf{x}$, generates an $N\,x\,1$ vector of wavelet-domain transform coefficients, $\mathbf{y}$. The $N\,x\,1$ vector $\mathbf{y}$ can be separated into two $\frac{N}{2}\,x\,1$ vectors of approximation and detail coefficients, $\mathbf{y}_{lp}$ and $\mathbf{y}_{hp}$, respectively. The spectral content of the signal $\mathbf{x}$ captured in $\mathbf{y}_{lp}$ and $\mathbf{y}_{hp}$ corresponds to the frequency subbands realized in 2:1 decimated output sequences from a QMF filterbank which obeys the "power complimentary condition", i.e.,

$$\left|H_{lp}(\theta)\right|^2 + \left|H_{lp}(\theta+\pi)\right|^2 = 1 \qquad (45)$$

where $H_{lp}(\theta)$ is the frequency response of the lowpass filter.

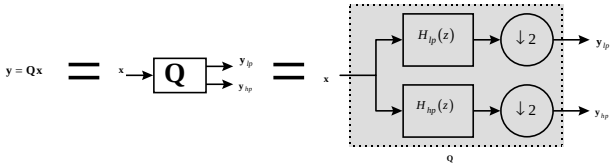


Fig. 15. Filterbank Interpretation of the DWT

Therefore, successive applications of the DWT can be interpreted as passing input data through a cascade of banks of perfect reconstruction lowpass (LP) and highpass (HP) filters followed by 2:1 decimation. In effect, the forward/inverse transform matrices of a particular wavelet are associated with a corresponding QMF analysis/synthesis filterbank. The usual wavelet decomposition implements an octave-band filterbank structure shown in Fig. 16. In the figure, frequency subbands associated with the coefficients from each stage are schematically represented for an audio signal sampled at 44.1 kHz.

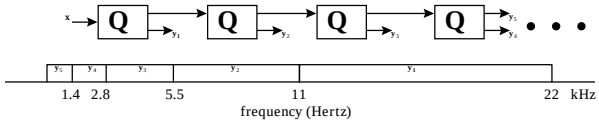


Fig. 16. Wavelet Decomposition

Wavelet packet representations, on the other hand, decompose both the detail and approximation coefficients at each stage of the tree, as shown in Fig. 17. In the figure, frequency subbands associated with the coefficients from each stage are schematically represented for an audio signal sampled at 44.1 kHz.

A filterbank interpretation of wavelet transforms is attractive in the context of audio coding algorithms for at least two reasons. First, wavelet or wavelet packet decompositions can be tree structured as necessary (unbalanced trees are possible) to decompose input audio into a set of frequency subbands tailored to some application. It is possible, for example, to approximate the critical band auditory filterbank utilizing a wavelet packet approach. Second, many K-coefficient finite

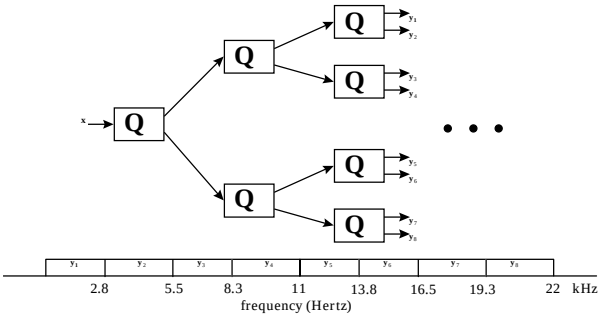


Fig. 17. Wavelet Packet Decomposition

support wavelets are associated with a single magnitude frequency response QMF pair, therefore a specific subband decomposition can be realized while retaining the freedom to choose a wavelet basis which is in some sense "optimal." For these and other reasons, several DWT-based subband algorithms for high-fidelity audio coding have been recently proposed.

The basic idea behind DWT-based subband coders is to quantize and encode efficiently the coefficient sequences associated with each stage of the wavelet decomposition tree. Irrelevancy is exploited by transforming frequency-domain masking thresholds to the wavelet domain and shaping wavelet-domain quantization noise such that it does not exceed the masking threshold. Wavelet-based subband algorithms also exploit statistical signal redundancies through differential, run-length, and entropy coding schemes. The next few subsections concentrate upon DWT-based subband coders developed by Tewfik *et al.* [71][72][73] during the last few years, including a very recently proposed hybrid sinusoidal/wavelet transform algorithm [74]. Other studies of DWT-based audio coding schemes concerned with low-complexity, low-delay, combined wavelet/multipulse LPC coding, and combined scalar/vector quantization of transform coefficients were reported, respectively, by Black and Zeytinoglu [65], Kudumakis and Sandler [66][67][68], Boland and Deriche [69], and Boland and Deriche [70].

D. ADAPTED WAVELET DECOMPOSITIONS

Sinha and Tewfik developed a variable-rate wavelet-based coding scheme for which they reported nearly transparent coding of CD-quality audio at 48-64 kbps [71][72]. The encoder (Fig. 18) exploits redundancy using a VQ scheme and irrelevancy using a wavelet packet (WP) signal decomposition combined with perceptual masking thresholds. The algorithm works as follows. Input audio is segmented into Nx1 vectors which are then windowed using a 1/16-th overlap square root Hann window. The dynamic dictionary (DD), which is essentially an adaptive VQ subsystem,

then eliminates signal redundancy. A dictionary of Nx1 codewords is searched for the vector which is perceptually closest to the input vector. The effective size of the dictionary is made larger than its actual size by a novel correlation lag search/time-warping procedure which identifies two N/2-sample codewords for each N-sample input vector. At both the transmitter and receiver, the dictionary is systematically updated with N-sample reconstructed output audio vectors according to a perceptual distance criterion and last-used-first-out rule. For irrelevancy reduction, an optimized WP decomposition is applied to the original signal as well as the DD residual. The decomposition tree is structured such that its 29 frequency subbands roughly correspond to the critical bands of the auditory filterbank. Psychoacoustic masking thresholds are derived using the same procedure as [61] and transformed to the wavelet domain so that WP coefficients can be quantized and encoded without introducing perceptible artifacts. Masking thresholds are assumed constant within each subband. The encoder transmits the particular combination of DD and WP information which minimizes the bit rate while maintaining perceptual quality. Three combinations are possible. In one scenario, the DD index and time-warping factor are transmitted alone if the DD residual energy is below the masking threshold at all frequencies. Alternatively, if the DD residual has audible noise energy, then WP coefficients of the DD residual are also quantized, encoded, and transmitted. In some cases, however, WP coefficients corresponding to the original signal are more compactly represented than the combination of the DD plus WP residual information. In this case, the DD information is discarded and only quantized and encoded WP coefficients are transmitted. In the latter two cases, the encoder also transmits subband scale factors, bit allocations, and energy normalization side information.

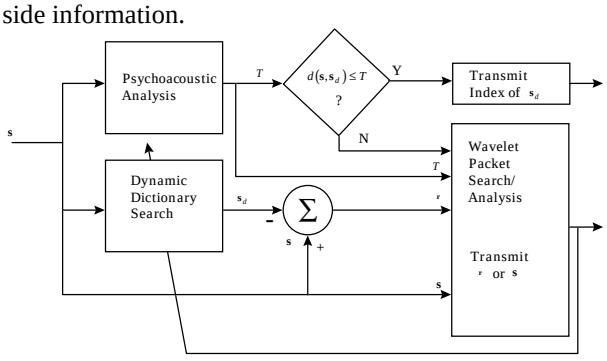


Fig. 18. Dynamic Dictionary/Optimal Wavelet Packet Encoder (after [71])

This algorithm is unique in that it contains the first reported application of adapted WP analysis to subband coding of high-fidelity, CD-quality audio. During each analysis frame, the WP basis selection procedure applies an optimality criterion of minimum bit rate for a given distortion level. The authors reached several useful conclusions regarding the choice of an optimal compact support (K-coefficient) wavelet basis. First, they found that basis optimization is warranted. Optimization produced average bit rate savings of 3, 6.5, 8.75, and 15 percent for wavelets selected from the sets associated with coefficient sequences of lengths 10, 20, 40, and 60, respectively. In an extreme case, a savings of 1.7 bits/sample is realized for transparent coding of a difficult castanets sequence when using best-case rather than worst-case wavelets (0.8 vs. 2.5 bits/sample for $K = 40$). Second, they determined that it is not necessary to search exhaustively the space of all wavelets for a particular value of K. The search can be limited to wavelets having the maximum possible number, or $\frac{K}{2}$ vanishing moments. The frequency responses of the filters associated with a $p-th$ order vanishing moment wavelet have $p-th$-order zeros at radian frequency $\theta = \pi$. Only a 3.1% bitrate reduction was realized for an exhaustive search versus a maximal vanishing moment constrained search. Third, the authors found that wavelets with longer coefficient sequences (larger K) tended to produce better results under the optimality constraint. Given identical distortion criteria for a castanets sequence, bit rates of 2.1 bits/sample for $K = 4$ wavelets were realized versus 0.8 bits/sample for $K = 40$ wavelets. Finally, deeper decomposition trees tended to yield better results, but the improvements saturated beyond a certain point.

As far as quality is concerned, subjective tests conducted by the authors with nine test subjects led them to conclude that the algorithm produced transparent quality for test material including drums, pop, violin with orchestra, and clarinet. Subjects detected differences between coded and original material, however, in the cases of castanets and piano sequences. The difficulty with castanets arises because of inadequate pre-echo control. This version of the coder utilizes only an adaptive window scheme which switches between 1024 and 2048-sample windows. Shorter windows (N=1024 or 23 ms) are used for signals which are likely to produce pre-echos. The piano sequence contained long segments of nearly steady or slowly decaying sinusoids. The wavelet coder does not handle steady sinusoids as well as other signals. With the exception of these troublesome signals in a comparative test, an additional expert listener also found that the WP coder outperformed MPEG-1, Layer II at 64 kbps.

Tewfik and Ali later enhanced the WP coder to improve pre-echo control and increase coding efficiency. After elimination of the dynamic dictionary, they reported better quality in the range of 55 to 63 kbps, as well as a real-time implementation of a simplified 64 to 78 kbps coder on two TMS320C31 devices [73]. Beyond DD removal, the major improvements included

exploitation of auditory temporal masking for pre-echo control, more efficient quantization and encoding of scale-factors, and run-length coding of long zero sequences. The improved coder also upgraded its psychoacoustic analysis section with a more sophisticated model similar to Johnston's PXFM coder [6]. The most notable improvement occurred in the area of pre-echo control. This was accomplished in the following manner. First, input frames likely to produce pre-echos are identified using a normalized energy measure criterion. These frames are parsed into 5 ms time slots (256 samples). Then, WP coefficients from all scales within each time slot are combined to estimate subframe energies. Masking thresholds computed over the global 1024-sample frame are assumed only to apply during high-energy time slots. Masking thresholds are reduced across all subbands for low energy time slots utilizing weighting factors proportional to the energy ratio between high- and low-energy time-slots. The remaining enhancements of improved scalefactor coding efficiency and run-length coding of zero sequences more than compensated for removal of the dynamic dictionary.

E. HYBRID HARMONIC/WAVELET DECOMPOSITION

Although the WP coder improvements reported in [73] addressed pre-echo control problems, they did not rectify the coder's inadequate performance for harmonic signals such as the piano test sequence. This is in part because wavelets do not provide compact representations for sinusoidal signals. On the other hand, wavelet decomposition techniques do provide signal representations that can efficiently track transients. Recognizing these facts, Hamdy *et al.* developed a novel hybrid coder [74] designed to exploit the efficiencies of both harmonic and wavelet signal representations. For each analysis frame, the encoder (Fig. 19) chooses a compact signal representation from combined sinusoidal and wavelet bases. This algorithm is based on the notion that short-time audio signals can be decomposed into tonal, transient, and noise components. It assumes that tonal components are most compactly represented in terms of sinusoidal basis functions, while transient and noise components are most efficiently represented in terms of wavelet bases. The encoder works as follows. First, Thompson's analysis model [75] is applied to extract sinusoidal frequencies, phases, and amplitudes for each input frame. Harmonic synthesis using the McAulay and Quatieri reconstruction algorithm [76] for phase and amplitude interpolation is next applied to obtain a residual sequence. Then, the residual is decomposed into WP subbands. The overall WP analysis tree approximates an auditory filterbank. Edge-detection processing identifies and removes transients in low frequency subbands. Once transients are eliminated, the residual WP coefficient sequences at each scale become largely decorrelated. In fact, the authors determined that the sequences are well approximated by white gaussian noise (WGN) sources having exponential decay envelopes. As far as quantization and encoding of the various parameters is concerned, sinusoidal frequencies are quantized with sufficient precision to satisfy psychoacoustic just-noticeable-differences in frequency (JNDF), which requires 8 bit absolute coding for a new frequency track, then 5 bit differential coding for the duration of the lifetime of the track. Sinusoidal amplitudes are quantized and encoded in a similar absolute/differential manner using simultaneous masking thresholds for shaping of quantization noise. This may require up to 8 bits per component. Sinusoidal phases are uniformly quantized on the interval $[-\pi, \pi]$ and encoded using 6-bits. As for quantization and encoding of WP parameters, all coefficients below 11 kHz are encoded as in [108]. Above 11 kHz, however, parametric representations are utilized. Transients are represented in terms of a binary edge mask which is then run-length encoded. Noise components are represented in terms of gaussian means, variances, and constants of exponential decay. The hybrid coder was reported to achieve nearly transparent coding over of wide range of CD-quality source material at bit rates in the vicinity of 44 kbps.

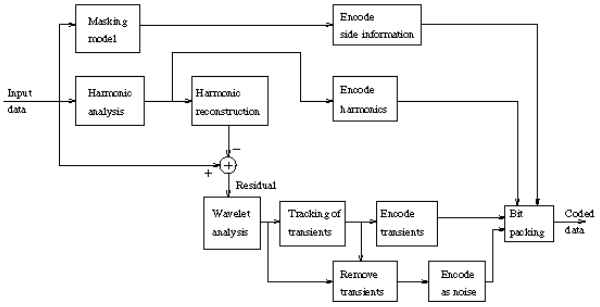
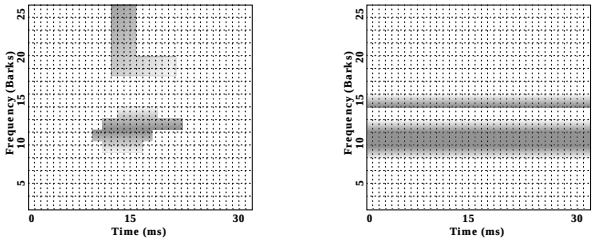


Fig. 19. Hybrid Sinusoidal/Wavelet Encoder (after [74])

F. SIGNAL-ADAPTIVE, NON-UNIFORM FILTERBANK (NUFB) DECOMPOSITIONS

This subsection introduces subband coding algorithms which utilize signal-adaptive banks of non-uniform bandpass filters to estimate the distribution of signal energy and masking power with respect to both time and frequency. The advantage of these systems over previous schemes is in superior time or frequency resolution matching between the analysis filterbank and the auditory filterbank. The disadvantage is that adaptive non-uniform filterbank design and implementation strategies are non-trivial. As we have seen, audio coding algorithms seek to achieve high coding gain by exploiting time-frequency signal decompositions which imitate the auditory filterbank. This task is difficult because of the non-uniform nature of critical bandwidth. For the purposes of time-frequency mapping, the algo-

rithms described in the previous three subsections make use of unitary transforms, uniform-resolution frequency subbands, and discrete wavelet or wavelet packet decompositions, respectively. Each type of algorithm makes some tradeoff between time resolution and frequency resolution. Transform coders typically offer very high frequency resolution at the expense of limited time resolution. The uniform subband algorithms, on the other hand, tend to offer good time resolution at the expense of frequency resolution. No resolution tradeoff is optimal for all audio signals. This dilemma is illustrated in Fig. 20 utilizing schematic representations of masking thresholds with respect to time and frequency for (a) a castanets and (b) a piccolo. In the figures, darker regions correspond to higher masking thresholds. For maximum coding gain, the strongly harmonic piccolo signal clearly calls for fine frequency resolution/coarse time resolution, because the masking thresholds remain relatively constant with respect time in several narrow frequency bands. Time resolution is not a big issue for this signal. Quite the opposite is true in the case of the castanets signal, however. Due to the fast attacks which characterize this percussive sound, masking thresholds are highly time-dependent in the 30 ms analysis window. Frequency resolution is not as important here because the thresholds tend to remain constant across a wide band of upper frequencies. All of this implies that an ideal coder should make adaptive decisions regarding optimal time-frequency signal decomposition.



(a) (b)

Fig. 20. Masking Thresholds in the Time-Frequency Plane: (a) Castanets, (b) Piccolo (after [80])

The most popular method for realizing nonuniform frequency subbands is to cascade uniform filters in an unbalanced tree structure. For a given impulse response length, however, cascade structures in general produce poor channel isolation. Recent advances in modulated filterbank design methodologies (e.g., [77]) have made tractable direct form near perfect reconstruction non-uniform designs which are critically sampled. This subsection describes coders which employ signal-adaptive non-uniform modulated filterbanks to approximate the time-frequency analysis properties of the auditory system more effectively than the uniform resolution algorithms described in prior sections. The discussion includes proposals by Sinha, Princen, and Johnston [80][81], as well as Purat and Noll [83][107]. In work not discussed here, other investigators have developed non-uniform filterbank coding techniques which address redundancy reduction utilizing lattice [78] and bidimensional VQ schemes [79].

G. AT&T SWITCHED NON-UNIFORM FILTERBANK CASCADE

Princen and Johnston developed a CD-quality coder based upon a signal-adaptive filterbank [80] for which they reported quality better than the sophisticated MPEG-1 Layer III algorithm at both 48 and 64 kbps. The analysis filterbank for this coder consists of a two-stage cascade. The first stage is a 48-band non-uniform modulated filterbank split into four uniform-bandwidth sections. There are eight uniform subbands from 0-750 Hz, four uniform subbands from 750-1500 Hz, 12 uniform subbands from 1.5-6 kHz, and 24 uniform subbands from 6-24 kHz. The second stage in the cascade optionally decomposes non-uniform bank outputs with on/off switchable banks of finer resolution uniform subbands. During filterbank adaptation, a suitable overall time-frequency resolution is attained by selectively enabling or disabling the second stage filters for each of the four uniform bandwidth sections. Low resolution mode for this architecture corresponds to slightly better than auditory filterbank frequency resolution. On the other hand, high-resolution mode for this architecture corresponds roughly to 512 uniform subband decomposition. Adaptation decisions are made independently for each of the four cascaded sections based on a criterion of minimum perceptual entropy (PE). The second stage filters in each section are enabled only if a reduction in PE (hence bit rate) is realized. Uniform PCM is applied to subband samples under the constraint of perceptually masked quantization noise. Masking thresholds are transmitted as side information. Further redundancy reduction is achieved by Huffman coding of both quantized subband sequences and masking thresholds. In informal listening tests, quality was reported to be better than the MPEG-1, Layer III coder at both 48 and 64 kbps.

H. AT&T SWITCHED MDCT/WAVELET FILTERBANK

Sinha and Johnston at Bell Labs later developed a different signal-adaptive switched filterbank coding scheme which achieved transparent coding of *stereo* CD-quality source material at 64 kbps per stereo pair [81]. Like the Princen and Johnston coder (above), this algorithm seeks to match the time-frequency distribution of masking power in the input signal with an appropriate analysis filterbank. Also as in [80], the filterbank switching criterion is a function of minimum PE. In contrast to the elaborate adaptive non-uniform filterbank cascade of Princen and Johnston's coder, however, this signal-adaptive algorithm switches between two distinct filterbanks. A 1024-point MDCT (Eq. 41) decomposition is applied normally, during "stationary"

periods. The coder switches to a tree-structured WP decomposition matched to the auditory filterbank during sharp transients. As noted previously [43][48], the MDCT lends itself to compact representation of stationary signals, and a 1024-point block size yields sufficiently high frequency resolution (Fig. 20b). The switch to a WP analysis during transients is warranted in light of the higher time resolution required at high frequencies for accurate estimation of the time/frequency distribution of masking power associated with sharp attacks (Fig. 20a). WP analysis also leads to more compact signal representation during transient periods. This coder offers the advantage of lower complexity than [80]. It also has an advantage over window length switching schemes [12] in that the desired improvement in time resolution during transient periods is restricted to the high frequency regions of interest. As for implementation details, the coder makes a switching decision every 25 ms. Carefully designed start and stop windows are inserted between analysis frames during switching intervals to mitigate boundary effects associated with the MDCT-to-WP transitions. Masking thresholds are estimated as in [6] and [8]. In subjective tests involving 12 expert and non-expert listeners with difficult castanets and triangle test signals, the coder outperformed the AT&T PAC algorithm [82] at 64 kbps per stereo pair by an average of 0.4-0.6 on a five-point quality scale.

I. FV-MLT

Purat and Noll [107] also developed a CD-quality audio coding scheme based on a signal-adaptive, nonuniform, tree-structured wavelet packet decomposition. This coder is unique in two ways. First of all, it makes use of a novel wavelet packet decomposition proposed in [83]. Secondly, the algorithm adapts to the signal the wavelet packet tree decomposition depth and breadth (branching structure) based on a minimum bit rate criterion, subject to the constraint of inaudible coefficient distortions. This in contrast to [81], which selects between two fixed filterbanks based on input signal characteristics. It also differs from [71], which applies a fixed wavelet packet decomposition tree structure but adapts the analysis wavelet. In informal subjective tests, the algorithm achieved excellent quality at a bit rate of 55 kbps.

V. AUDIO CODING STANDARDS

This section gives high-level descriptions of some international and commercial product audio coding standards, including the ISO/IEC MPEG-1/-2 series, the Sony ATRAC, the Phillips DCC, and the Dolby AC-3 algorithms.

A. ISO/IEC 11172-3 (MPEG-1) AND ISO/IEC IS13818-3 (MPEG-2)

An International Standards Organization/Moving Pictures Experts Group (ISO/MPEG) audio coding standard for stereo CD-quality audio was adopted in 1992 after four years of extensive collaborative research by audio coding experts worldwide. ISO 11172-3 [84] comprises a flexible hybrid coding technique which incorporates several methods including subband decomposition, filterbank analysis, transform coding, entropy coding, dynamic bit allocation, nonuniform quantization, adaptive segmentation, and psychoacoustic analysis. MPEG coders accept 16-bit PCM input data at sample rates of 32, 44.1, and 48 kHz. MPEG-1 (1992) offers separate modes for mono, stereo, dual independent mono, and joint stereo. Available bit rates are 32-192 kb/s for mono and 64-384 kb/s for stereo. MPEG-2 (1994) [85][86][87] extends the capabilities offered by MPEG-1 to support the so called 3/2 channel format with left, right, center, and left and right surround channels. The first MPEG-2 standard is backward compatible with MPEG-1 in the sense that 3/2 channel information transmitted by an MPEG-2 encoder can be correctly decoded for 2-channel presentation by an MPEG-1 receiver. The second MPEG-2 standard sacrifices backwards MPEG-1 compatibility to eliminate quantization noise unmasking artifacts [88] which are potentially introduced by the forced backward compatibility. Several discussions of the MPEG-1 [89] and MPEG-1/2 [26][27] standards have appeared recently in the literature. MPEG standardization work is continuing and will eventually lead to very low rate high fidelity coding, perhaps reaching bit rates as low as 24 kb/s per channel.

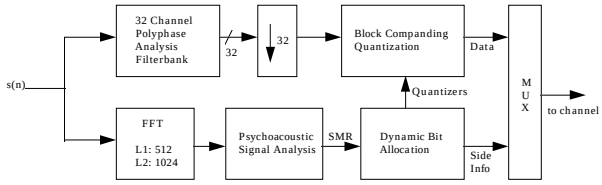


Fig. 21. ISO/MPEG Layer I/II Encoder

The MPEG-1 architecture contains three layers of increasing complexity, delay, and output quality. Each higher layer incorporates functional blocks from the lower layers. Layers I and II (Fig. 21) work as follows. The input signal is first decomposed into 32 critically sub-sampled subbands using a polyphase filterbank. These 511th-order filters are equally spaced such that a 48 kHz input signal is split into 750 Hz subbands, with the subbands decimated 32:1. In the absence of quantization noise, each filter would perfectly cancel aliasing introduced by adjacent bands. In practice, however, the filters are designed for very high sidelobe attenuation (96 dB) to insure that intra-band aliasing due to quantization noise remains negligible. For the purposes of psychoacoustic analysis and determination of JND thresholds, a 512 (layer I) or 1024 (layer II) point FFT is computed in parallel with the subband decomposition for each decimated block of 12 input samples (8 ms at

48 kHz). Next, the subbands are block companded (normalized by a scalefactor) such that the maximum sample amplitude in each block is unity, then an iterative bit allocation procedure applies the JND thresholds to select an optimal quantizer from a predetermined set for each subband. Quantizers are selected such that both the masking and bit rate requirements are simultaneously satisfied. In each subband, scalefactors are quantized using 6 bits and quantizer selections are encoded using 4 bits. For layer I encoding, decimated subband sequences are quantized and transmitted to the receiver in conjunction with side information, including quantized scalefactors and quantizer selections. With operation similar to layer I, layer II offers enhanced output quality and reduced bit rates at the expense of greater complexity and increased delay. Improvements occur in three areas. First, the psychoacoustic threshold determination benefits from better frequency resolution because of the increased FFT size. Second, scalefactor side information is reduced by considering properties of three adjacent 12-sample blocks and optionally transmitting one, two, or three scalefactors as well a 2-bit side parameter to indicate how many are being transmitted (increases delay). Third, maximum subband quantizer resolution is increased to 16 bits from the layer I limit of 15. The overall bit rate is reduced in spite of this increase by decreasing the number of available quantizers with increasing subband index. Average Mean Opinion Scores (MOS) of 4.7 and 4.8 have been reported [26] for monophonic layer I and layer II codecs operating at 192 and 128 kb/s, respectively. Averages were computed over a range of test material.

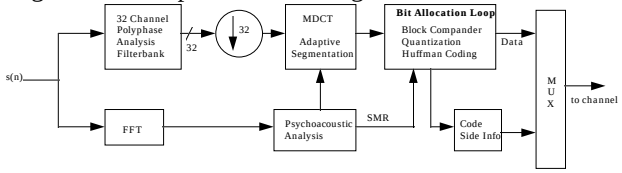


Fig. 22. ISO/MPEG Layer III Encoder

The layer III MPEG (Fig. 22) architecture achieves performance improvements by adding several important mechanisms on top of the layer I/II foundation. A hybrid filterbank is introduced to increase frequency resolution and thereby better approximate critical band behavior. The hybrid filterbank includes adaptive segmentation to improve pre-echo control. Sophisticated bit allocation and quantization strategies which rely upon non-uniform quantization, analysis-by-synthesis, and entropy coding are introduced to allow reduced bit rates and improved quality. First, a hybrid filterbank is constructed by following each subband filter with an adaptive MDCT. This practice allows for higher frequency resolution and pre-echo control. Use of an 18-point MDCT, for example, improves frequency resolution to 41.67 Hz per spectral line. Adaptive MDCT block sizes between 6 and 18 points to allow improved pre-echo control. Using shorter blocks during rapid attacks in the input sequence (6 decimated points at 48 kHz = 4 ms) allows premasking to hide pre-echoes, while using longer blocks during steady-state periods reduces side information and hence bit rates. Bit allocation and quantization of the spectral lines is realized in a nested loop procedure which uses both non-uniform quantization and Huffman coding. The inner loop adjusts the non-uniform quantizer step sizes for each block until the number of bits required to encode the transform components falls within the bit budget. The outer loop evaluates the quality of the coded signal (analysis-by-synthesis) in terms of quantization noise relative to the JND thresholds. Average MOS of 3.1 and 3.7 were reported [26] for monophonic layer II and layer III codecs operating at 64 kbps.

B. PRECISION ADAPTIVE SUBBAND CODING (PASC)

Phillips Digital Compact Cassette (DCC) is an example of a consumer product which essentially implements the 384 kb/s stereo mode of MPEG-1, layer I. A discussion of the Precision Adaptive Subband Coding algorithm and other elements of the DCC system are given in [90].

C. ADAPTIVE TRANSFORM ACOUSTIC CODING (ATRAC)

The ATRAC coding method developed by Sony for use in its rewritable MiniDisc system makes combined use of subband and transform coding techniques to achieve CD-quality 256 kb/s coding of 44.1 kHz stereo 16-bit PCM input data [91]. The ATRAC encoder first splits the input signal into three subbands which cover the ranges of 0-5.5 kHz, 5.5-11 kHz, and 11-22 kHz using a QMF analysis filterbank. Like MPEG layer III, the ATRAC QMF filterbank is followed by adaptive MDCT analysis. Although long block sizes of 11.6 ms are normally selected for all subbands, short block sizes of 1.45 ms in the high-frequency band and 2.9 ms in the low and mid-frequency bands are used to affect pre-echo cancellation during input attack periods. Finally, MDCT components are quantized and encoded according to a psychoacoustically derived bit allocation.

D. DOLBY AC-3

Dolby Laboratories originally developed the 320 kb/s AC-3 perceptual audio coder [92] for High-Definition Television (HDTV). Its first application, however, has been in the cinema. Digital information is interleaved between sprocket holes on one side of the 35 mm film. The coder carries 5.1 channels of audio (left, center, right, left surround, right surround, and a subwoofer), but it has also been designed for compatibility with conventional mono, stereo, and matrixed multi-channel sound reproduction systems. The PCM input signal is first windowed using a proprietary function and then segmented into 50% overlapping 10.66 ms blocks (512 samples). The block size is reduced to 5.33

ms during transient conditions to compensate for pre-echos. After segmentation, a modified Discrete Cosine Transform (MDCT) filterbank with 93.75 Hz frequency resolution is used to decompose the signal. The MDCT offers a good compromise between frequency resolution and time resolution, since it is critically sampled with 50% time overlap. Transform components are quantized using a psychoacoustically derived dynamic bit allocation scheme. Spectral information obtained from the MDCT is encoded using a novel mantissa/exponent coding scheme as follows. First, the spectral stability is evaluated. All transform coefficients are transmitted for stable spectra, but time updates occur only every 32 ms. Fewer components are encoded for transient signals, but time updates occur frequently, e.g., every 5.3 ms. A spectral envelope is formed from exponents corresponding to log spectral line magnitudes. These exponents are differentially encoded. Psychoacoustic quantization masking thresholds are derived from the decoded spectral envelope for 64 non-uniform subbands which increase in size proportional to the ear's critical bands. The thresholds are used to select appropriate quantizers for transform coefficient mantissas in a bit allocation loop. If too few bits are available, high-frequency coupling (above 2 kHz) between channels may be used to reduce the amount of transmitted information. Exponents, mantissas, coupling data, and exponent strategy data are combined and transmitted to the receiver. AC-3 has been selected for use in the United States HDTV system [97]. It is also being designed into other consumer electronics equipment such as cable television and direct broadcast satellite.

VI. CONCLUSION

A. SUMMARY OF APPLICATIONS FOR COMMERCIAL AND INTERNATIONAL STANDARDS

Current applications (Table 2) which benefit from audio coding include digital broadcast audio (DBA) [93][94], Direct Broadcast Satellite (DBS) [95], Digital Versatile Disk (DVD) [96], high-definition television (HDTV) [97], cinematic theater [98], and audio-on-demand over wide area networks such as the Internet [99]. Audio coding has also enabled miniaturization of digital audio storage media such as Compact MiniDisk [100] and Digital Compact Cassette (DCC) [101][102].

B. SUMMARY OF RECENT RESEARCH

The level of sophistication and high performance achieved by the standards listed in Table 2 reflects the fact that audio coding algorithms have matured rapidly in less than a decade. The emphasis nowadays has shifted to realizations of low-rate, low-complexity, and low-delay algorithms [103]. Using primarily transform [104][105][106], subband (filterbank/wavelet) [107] [108][109][110][111], and other [112][113][114] coding methodologies coupled with perceptual bit allocation strategies, new algorithms continue to advance the state-of-the art in terms of bit rates and quality (Table 3). Sinha and Johnston, for example, reported transparent CD-quality at 64/32 kbps for stereo/mono [110] sources. Other new algorithms include extended capacity for multi-channel/ multi-language systems [98][115] [116].

Algorithm	**Method**	**Sample Rates (kHz)**	**Chan.**	**References**
APT-X100	ADPCM	44.1	1	
ATRAC	Subband	44.1	2	[100]
AT&T PAC	Trans.	44.1	1 - 5.1	[16]
Dolby AC-2	Trans.	44.1	2	[2]
Dolby AC-3	Trans.	44.1	1 - 5.1	[98]
MPEG-1 Layers I-III	Hybrid	32,44.1,48	1,2	[17]
MPEG-2 Layers I-III	Hybrid	16,22,24 32,44.1,48	1 - 5.1	[18]

Algorithm	**Bit Rate(s) (kbps)**	**Quality**	**Applications**
APT-X100	176,400	CD	Cinema
ATRAC	256/ch	CD	MiniDisc
AT&T PAC	128/stereo	FM,CD	DBA: 128/160 kbps
Dolby AC-2	256/ch	CD	DBA
Dolby AC-3	32-384	CD	Cinema, HDTV
MPEG-1 Layers I-III	32-448	AM,FM, CD	DBA: LII@256 kbps DBS: LII@224 kbps DCC: LI@384 kbps
MPEG-2 Layers I-III	16-	AM,FM CD	Misc. network

Table 2. Audio Coding Standards and Applications

Methodology	**Bit Rate(s) (kbps)**	**References**
Signal adaptive switched filterbank	below 32	Sinha,Johnston,Princen [80][81]
Sinusoidal/wavelet packet	44	Sinha, Tewfik [74]
Adapted wavelet packet	48-63	Sinha, Tewfik [72]
Frequency varying modulated lapped transforms	55	Purat, Noll [83],[107]
Transform/subband noise substitution	56	Schulz [50]
Transform/differential frame encoding (DPAC)	60-100	Paraskevas, Mourjopoulos [106]
Wavelet packet/ multipulse LPC	64	Boland,Deriche[69]
Transform domain weighted interleave VQ	below 64	Iwakami, et al., [53],[105]
Modulated phasor transform/bidimensional lattice	80	Mainard, Lever [111]
Nonuniform filterbank/ lattice VQ	96	Monta, Cheung [78]

Table 3. Recent Audio Coding Research

C. FUTURE DIRECTIONS

In addition to the pursuing the usual goals of transparent compression at lower bit rates (below 64 kbps/channel) with reduced complexity, minimal delay, and enhanced bit error robustness, future research in audio coding will be concerned with the development of algorithms which offer scalability [117][118][119]. This trend is reflected in the set of candidate algorithms [120] proposed for MPEG-4 standardization [121], as shown in Table 4.

Company/ Institution	Methodology	Bit Rates (kbps)	Scalability Modes
Alcatel / Philips / RAI	subband, ADPCM, adaptive Huffman	24, 40	64/6 64/24/6
AT&T	transform	24, 40	64/6
AT&T	waveform interp.	2, 6	-
Bosch / CSELT / MATRA	subband/LPC, fine step scalability	6	64/24/6
INESC	subband/transform, Huffman, harmonic component extraction	16, 24, 40	-
Matsushita	CELP + post-processing for speed control + enhancement for error robustness	6	-
Motorola	transform, Huffman,enhancement for error robustness	24, 40	64/24/6
NEC	transform, enhancement for scalability	24, 40	1,2
NEC	CELP with multimode coding	6	1,2
NTT	transform, VQ, enhancement for error robustness	16,24,40	2
NTT	CELP, pitch synchronous innovation, enhancement for error robustness	2, 6	3
NTT DoCoMo	CELP, pitch synchronous innovation, enhancement for error robustness	-	
Philips	CELP with efficient search	16	-
Samsung	subband, VQ	-	1
Sony	subband/trans., scalability with LPC	-	1
Sony IPC	subband/trans., LPC residual coding, enhancement for scalability	-	3
Sony IPC	LPC with harmonic vector excitation, enhancement for error robustness	2, 6	3
University of Erlangen	transform, scalability with low bit rate codecs	24, 40	1,2
University of Hannover / Deutsche Telekom	analysis/synthesis coding for individual spectral lines + enhancement for scalability	6, 16	-
JVC	transform, VQ	40	-

Table 4. Candidate Codecs Proposed for MPEG-4 Standardization (after [122])

References

1. International Electrotechnical Commission/ American National Standards Institute (IEC/ANSI) CEI-IEC-908, "Compact Disc Digital Audio System" ("red book"), 1987.
2. C. Todd, "A Digital Audio System for Broadcast and Prerecorded Media," in *Proc. 75th Conv. Aud. Eng. Soc.*, preprint #, Mar. 1984.
3. E.F. Schroder and W. Voessing, "High Quality Digital Audio Encoding with 3.0 Bits/Sample using Adaptive Transform Coding," in *Proc. 80th Conv. Aud. Eng. Soc.*, preprint #2321, Mar. 1986.
4. G. Theile, *et al.*, "Low-Bit Rate Coding of High Quality Audio Signals," in *Proc. 82nd Conv. Aud. Eng. Soc.*, preprint #2432, Mar. 1987.
5. K. Brandenburg, "OCF - A New Coding Algorithm for High Quality Sound Signals," in *Proc. ICASSP-87*, pp. 5.1.1-5.1.4, May 1987.
6. J. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," *IEEE J. Sel. Areas in Comm.*, pp. 314-323, Feb. 1988.
7. W-Y Chan and A. Gersho, "High Fidelity Audio Transform Coding with Vector Quantization," in *Proc. ICASSP-90*, pp. 1109-1112, May 1990.
8. K. Brandenburg and J.D. Johnston, "Second Generation Perceptual Audio Coding: The Hybrid Coder," in *Proc. 88th Conv. Aud. Eng. Soc.*, preprint #2937, Mar. 1990.
9. K. Brandenburg, et al., "ASPEC: Adaptive Spectral Entropy Coding of High Quality Music Signals," in Proc. 90th Conv. Aud. Eng. Soc., preprint #3011, Feb. 1991.

10. Y.F. Dehery, et al., “A MUSICAM Source Codec for Digital Audio Broadcasting and Storage,” in Proc. ICASSP-91, pp. 3605-3608, May 1991.
11. M. Iwadare, et al., “A 128 kb/s Hi-Fi Audio CODEC Based on Adaptive Transform Coding with Adaptive Block Size MDCT,” IEEE J. Sel. Areas in Comm., pp. 138-144, Jan. 1992.
12. K. Brandenburg et al., “ISO-MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio,” J. Audio Eng. Soc., pp. 780-792, Oct. 1994.
13. G. Stoll, et al., “Generic Architecture of the ISO/MPEG Audio Layer I and II: Compatible Developments to Improve the Quality and Addition of New Features,” in Proc. 95th Conv. Aud. Eng. Soc., preprint #3697, Oct. 1993.
14. J.B. Rault, et al., “MUSICAM (ISO/MPEG Audio) Very Low Bit-Rate Coding at Reduced Sampling Frequency,” in Proc. 95th Conv. Aud. Eng. Soc., preprint #3741, Oct. 1993.
15. G. Stoll, et al., Extension of ISO/MPEG-Audio Layer II to Multi-Channel Coding: The Future Standard for Broadcasting, Telecommunication, and Multilmedia Applications,” in Proc. 94th Conv. Aud. Eng. Soc., preprint #3550, Mar. 1993.
16. J.D. Johnston, et al., “The AT&T Perceptual Audio Coder (PAC),” Presented at the AES convention, New York, Oct., 1995.
17. ISO/IEC JTC1/SC29/WG11 MPEG, IS11172-3 “Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mbit/s, Part 3: Audio” 1992. (“MPEG-1”)
18. ISO/IEC JTC1/SC29/WG11 MPEG, IS13818-3 “Information Technology - Generic Coding of Moving Pictures and Associated Audio, Part 3: Audio” 1994. (“MPEG-2”)
19. F. Wylie, “Predictive or Perceptual Coding...apt-X and apt-Q,” in Proc. 100th Conv. Aud. Eng. Soc., preprint #4200, May 1996.
20. P. Craven and M. Gerzon, “Lossless Coding for Audio Discs,” J. Audio Eng. Soc., pp. 706-720, Sep. 1996.
21. J. R. Stuart, “A Proposal for the High-Quality Audio Application of High-Density CD Carriers,” Technical Subcommittee Acoustic Renaissance for Audio, http://www.meridian.co.uk/ara/araconta.html, pp. 1-26, Jun. 1995.
22. N. Jayant, et al., “Coding of Wideband Speech,” Speech Comm., pp. 127-138, Jun. 1992.
23. N. Jayant, “High Quality Coding of Telephone Speech and Wideband Audio,” in Advances in Speech Signal Processing, S. Furui and M.M. Sondhi, Eds., New York: Dekker, 1992.
24. J. Johnston and K. Brandenburg, “Wideband Coding - Perceptual Considerations for Speech and Music,” in Advances in Speech Signal Processing, S. Furui and M.M. Sondhi, Eds., New York: Dekker, 1992.
25. N. Jayant, et al., “Signal Compression Based on Models of Human Perception,” Proc. IEEE, pp. 1385-1422, Oct. 1993.
26. P. Noll, “Wideband Speech and Audio Coding,” IEEE Comm. Mag., pp.34-44, Nov. 1993.
27. P. Noll, “Digital Audio Coding for Visual Communications,” Proc. IEEE, pp. 925-943, Jun. 1995.
28. H. Fletcher, “Auditory Patterns,” Rev. Mod. Phys., pp. 47-65, Jan. 1940.
29. D.D. Greenwood, “Critical Bandwidth and the Frequency Coordinates of the Basilar Mem.brane,” J. Acous. Soc. Am., pp. 1344-1356, Oct. 1961.
30. J. Zwislocki, “Analysis of Some Auditory Characteristics,” in Handbook of Mathematical Psychology, R. Luce, et al., Eds., New York: John Wiley and Sons, Inc., 1965.
31. B. Scharf, “Critical Bands,” in Foundations of Modern Auditory Theory, New York: Academic Press, 1970.
32. R. Hellman, “Asymmetry of Masking Between Noise and Tone,” Percep. and Psychphys., pp. 241-246, vol.11, 1972.
33. E. Zwicker and H. Fastl, Psychoacoustics Facts and Models, Springer-Verlag, 1990.
34. E. Zwicker and U. Zwicker, “Audio Engineering and Psychoacoustics: Matching Signals to the Final Receiver, the Human Auditory System,” J. Audio Eng. Soc., pp. 115-126, Mar. 1991.
35. M. Schroeder, et al., “Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear,” J. Acoust. Soc. Am., pp. 1647-1652, Dec. 1979.
36. J. Johnston, “Estimation of Perceptual Entropy Using Noise Masking Criteria,” in Proc. ICASSP-88, pp. 2524-2527, May 1988.
37. Terhardt, E., “Calculating Virtual Pitch,” Hearing Research, pp. 155-182, 1, 1979.
38. N. Jayant, et al., “Signal Compression Based on Models of Human Perception,” Proc. IEEE, pp. 1385-1422, Oct. 1993.
39. P. Papamichalis, “MPEG Audio Compression: Algorithms and Implementation,” in Proc. DSP 95 Int. Conf. on DSP, pp. 72-77, June 1995.
40. N. Jayant and P. Noll, Digital Coding of Waveforms Principles and Applications to Speech and Video, Englewood Cliffs: Prentice-Hall, 1984.
41. D. Krahe, “New Source Coding Method for High Quality Digital Audio Signals,” NTG Fachtagung Hoerrundfunk, Mannheim, 1985.

42. D. Krahe, "Grundlagen eines Verfahrens zur Dataenreduktion bei Qualitativ Hochwertigen, digitalen Audiosignalen auf Basis einer Adaptiven Transformationscodierung unter Berucksightigung Psychoakustischer Phanomene," Ph.D. Thesis, Duisburg 1988.
43. K. Brandenburg, "High Quality Sound Coding at 2.5 Bits/Sample," *in Proc. 84th Conv. Aud. Eng. Soc.*, preprint #2582, Mar. 1988.
44. K. Brandenburg, "OCF: Coding High Quality Audio with Data Rates of 64 kbit/sec," *in Proc. 85th Conv. Aud. Eng. Soc.*, preprint #2723, Mar. 1988.
45 J. Johnston, "Perceptual Transform Coding of Wideband Stereo Signals," in *Proc. ICASSP-89*, pp. 1993-1996, May 1989.
46 J. Jetzt, "Critical Distance Measurements on Rooms from the Sound Energy Spectrum Response," *J. Acoust. Soc. Am.*, pp. 1204-1211, 1979.
47. Y. Mahieux, *et al.*, "Transform Coding of Audio Signals Using Correlation Between Successive Transform Blocks," in *Proc. Int. Conf. Acous., Speech, and Sig. Process.* (ICASSP-89), pp. 2021-2024, May 1989.
48. Y. Mahieux and J. Petit, "Transform Coding of Audio Signals at 64 kbits/sec," in *Proc. Globecom '90*, pp. 405.2.1-405.2.5, Nov. 1990.
49. A. Sugiyama, *et al.*, "Adaptive Transform Coding with an Adaptive Block Size (ATC-ABS)," in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-90), pp. 1093-1096, May 1990.
50. D. Schulz, "Improving Audio Codecs by Noise Substitution," *J. Audio Eng. Soc.*, pp. 593-598, Jul./Aug., 1996.
51. W. Chan and A. Gersho, "Constrained-Storage Vector Quantization in High Fidelity Audio Transform Coding," in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-91), pp. 3597-3600, May 1991.
52. W. Chan and A. Gersho, "Constrained-Storage Quantization of Multiple Vector Sources by Codebook Sharing," *IEEE Trans. Comm.*, Jan 1991.
53. T. Moriya, *et al.*, "Extension and Complexity Reduction of TWINVQ Audio Coder," in *Proc. Int. Conf. Acous., Speech, and Sig. Process.* (ICASSP-96), pp. 1029-1032, May 1996.
54. K. Ikeda, *et al.*, "Error Protected TwinVQ Audio Coding at Less Than 64 kbit/s," in *Proc. IEEE Speech Coding Workshop*, pp. 33-34, 1995.
55 J. Princen and A. Bradley, "Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation," *IEEE Trans. ASSP*, pp. 1153-1161, Oct. 1986.
56. P. Duhamel, *et al.*, "A Fast Algorithm for the Implementation of Filter Banks Based on Time Domain Aliasing Cancellation," in *Proc. Int. Conf. Acous., Speech, and Sig. Process.* (ICASSP-91), pp. 2209-2212, May 1991.
57. A. Charbonnier and J.P. Petit, "Sub-band ADPCM Coding for High Quality Audio Signals," in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-88), pp. 2540-2543, May 1988.
58. P. Voros, "High-Quality Sound Coding Within 2x64 kbit/s Using Instantaneous Dynamic Bit-Allocation," in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-88), pp. 2536-2539, May 1988.
59. D. Teh, *et al.*, "Subband Coding of High-Fidelity Quality Audio Signals at 128 kbps," in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-92), pp. II-197-II-200, May 1990.
60. G. Stoll, *et al.*, "Masking-Pattern Adapted Subband Coding: Use of the Dynamic Bit-Rate Margin," *in Proc. 84th Conv. Aud. Eng. Soc.*, preprint #2585, Mar. 1988.
61. R.N.J. Veldhuis, "Subband Coding of Digital Audio Signals without Loss of Quality," in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-89), pp. 2009-2012, May 1989.
62. D. Wiese and G. Stoll, "Bitrate Reduction of High Quality Audio Signals by Modelling the Ear's Masking Thresholds," *in Proc. 89th Conv. Aud. Eng. Soc.*, preprint #2970, Sep. 1990.
63. Swedish Broadcasting Corporation, "ISO MPEG/Audio Test Report," Stokholm, Jul. 1990.
64. I. Daubechies, *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, 1992.
65. M. Black and M. Zeytinoglu, "Computationally Efficient Wavelet Packet Coding of Wideband Stereo Audio Signals," in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-95), pp. 3075-3078, May 1995.
66. P. Kudumakis and M. Sandler, "On the Performance of Wavelets for Low Bit Rate Coding of Audio Signals," in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-95), pp. 3087-3090, May 1995.
67. P. Kudumakis and M. Sandler,"Wavelets, Regularity, Complexity, and MPEG-Audio," *in Proc. 99th Conv. Aud. Eng. Soc.*, preprint #4048, Oct. 1995.
68. P. Kudumakis and M. Sandler, "On the Compression Obtainable with Four-Tap Wavelets," *IEEE Sig. Proc. Let.*, pp. 231-233, Aug. 1996.
69. S. Boland and M. Deriche, "High Quality Audio Coding Using Multipulse LPC and Wavelet Decomposition," in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-95), pp. 3067-3069, May 1995.

70. S. Boland and M. Deriche, “Audio Coding Using the Wavelet Packet Transform and a Combined Scalar-Vector Quantization,” in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-96), pp. 1041-1044, May 1996.
71. D. Sinha and A. Tewfik, “Low Bit Rate Transparent Audio Compression Using a Dynamic Dictionary and Optimized Wavelets,” in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-93), pp. I-197-I-200, May 1993.
72. D. Sinha and A. Tewfik, “Low Bit Rate Transparent Audio Compression Using a Adapted Wavelets,” *IEEE Trans. Sig. Proc.*, pp. 3463-3479, Dec. 1993.
73. A. Tewfik and M. Ali, “Enhanced Wavelet Based Audio Coder,” in *Conf. Rec. of the 27th Asilomar Conf. on Sig. Sys., and Comp.*, pp. 896-900, Nov 1993.
74. K. Hamdy, *et al.*, “Low Bit Rate High Quality Audio Coding with Combined Harmonic and Wavelet Representations,” in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-96), pp. 1045-1048, May 1996.
75. D. Thompson, “Spectrum Estimation and Harmonic Analysis,” *Proc. IEEE*, pp. 1055-1096, Sep. 1982.
76. R. McAulay and T. Quatieri, “Speech Analysis Synthesis Based on a Sinusoidal Representation,” *IEEE Trans. ASSP*, pp. 744-754, Aug. 1986.
77. J. Princen, “The Design of Non-Uniform Modulated Filterbanks,” in *Proc. IEEE Int. Symp. on Time-Frequency and Time-Scale Analysis*, Oct. 1994.
78. P. Monta and S. Cheung, “Low Rate Audio Coder with Hierarchical Filterbanks,” in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-94), pp. II-209-II-212, May 1994.
79. L. Mainard and M. Lever, “A Bi-Dimensional Coding Scheme Applied to Audio Bit Rate Reduction,” in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-96), pp. 1017-1020, May 1994.
80. J. Princen and J. Johnston, “Audio Coding with Signal Adaptive Filterbanks,” in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-95), pp. 3071-3074, May 1995.
81. D. Sinha and J. Johnston, “Audio Compression at Low Bit Rates Using a Signal Adaptive Switched Filterbank,” in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-96), pp. 1053-1056, May 1996.
82. J. Johnston, “Sum-Difference Stereo Transform Coding,” in *Proc. Int. Conf. Acous., Speech, and Sig. Proc.* (ICASSP-92), pp. II-569-II-572, May 1992.
83. M. Purat and P. Noll, “A New Orthonormal Wavelet Packet Decomposition for Audio Coding Using Frequency-Varying Modulated Lapped Transforms,” in *IEEE ASSP Workshop on Applic. of Sig. Proc. to Aud. and Acous.*, Session 8, Oct. 1995.
84. ISO/IEC JTC1/SC29, “Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mbit/s - IS 11172-3 (audio),” 1992.
85. G. Stoll, *et al.*, “Extension of the ISO/MPEG-Audio Layer II to Multi-Channel Coding: The Future Standard for Broadcasting, Telecommunication, and Multimedia Application,” in *Proc. 94th Audio Eng. Soc. Conv.*, preprint 3550, Berlin, 1993.
86. B. Grill, *et al.*, “Improved MPEG-2 Audio Multi-Channel Encoding,” in *Proc. 96th Audio Eng. Soc. Conv.*, preprint 3865, Amsterdam, 1994.
87. ISO/IEC JTC1/SC29, “Information Technology - Generic Coding of Moving Pictures and Associated Audio Information - DIS 13818-3 (Audio),” 1994.
88. W. Th. ten Kate *et al.*, “Compatibility Matrixing of Multi-Channel Bit Rate Reduced Audio Signals,” in *Proc. 96th Audio Eng. Soc. Conv.*, perprint 3792, Amsterdam, 1994.
89. K. Brandenburg and G. Stoll, “ISO-MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio,” *J. Audio Eng. Soc.*, pp. 780-792, Oct. 1994.
90. A. Hoogendoorn, “Digital Compact Cassette,” *Proc. IEEE*, pp. 1479-1489.
91. T. Yoshida, “The Rewritable MiniDisc System,” *Proc. IEEE*, pp. 1492-1500, Oct. 1994.
92. C. Todd, et. al., “AC-3: Flexible Perceptual Coding for Audio Transmission and Storage,” in *Proc. 96th Conv. Aud. Eng. Soc.*, preprint #3796, Feb. 1994.
93. G. Stoll, “A Perceptual Coding Technique Offering the Best Compromise Between Quality, Bit-Rate, and Complexity for DSB,” in *Proc. 94th Audio Eng. Soc. Conv.*, preprint 3458, Berlin, 1993.
94. R. K. Jurgen, “Broadcasting with Digital Audio,” *IEEE Spectrum*, pp. 52-59, Mar. 1996.
95. Pritchard, “Direct Broadcast Satellite,” *Proc. IEEE*, pp. 1116-11, Jul. 1990.
96. P. Craven and M. Gerzon, “Lossless Coding for Audio Discs,” *J. Aud. Eng. Soc.*, pp. 706-720, Sep. 1996.
97. United States Advanced Television Systems Committee (ATSC), Audio Specialist Group (T3/S7) Doc. A/52, “Digital Audio Compression Standard (AC-3),” Nov. 1994.
98. C. Todd, *et al.*, “AC-3: Flexible Perceptual Coding for Audio Transmission and Storage,” in *Proc.*

96th Conv. Aud. Eng. Soc., preprint #3796, Feb. 1994.

99. M. Dietz, *et al.*, "Audio Compression for Network Transmission," *J. Audio Eng. Soc.*, pp. 58-70, Jan./Feb. 1996.
100. T. Yoshida, "The Rewritable MiniDisc System," *Proc. IEEE*, pp. 1492-1500, Oct. 1994.
101. G.C.P. Lokhoff, "Precision adaptive sub-band coding (PASC) for the digital compact cassette (DCC)," *IEEE Trans. Consumer Electron.*, pp. 784-789, Nov. 1992.
102. A. Hoogendoorn, "Digital Compact Cassette," *Proc. IEEE*, pp. 1479-1489, Oct. 1994.
103. ISO/IEC JTC1/SC29/WG11 MPEG94/443, "Requirements for Low Bitrate Audio Coding/MPEG-4 Audio", 1994. ("MPEG-4")
104. Y. Mahieux and J.P. Petit, "High-Quality Audio Transform Coding at 64 kbps," *IEEE Trans. Comm.*, pp. 3010-3019, Nov. 1994.
105. N. Iwakami, *et al.*, "High-Quality Audio-Coding at Less Than 64 kbit/s by Using Transform-Domain Weighted Interleave Vector Quantization (TWINVQ)," in *Proc. ICASSP-95*, pp. 3095-3098, May. 1995.
106. M. Paraskevas and J. Mourjopoulos, "A Differential Perceptual Audio Coding Method with Reduced Bitrate Requirements," *IEEE Trans. Speech and Audio Proc.*, pp. 490-503, Nov. 1995.
107. M. Purat and P. Noll, "Audio Coding with a Dynamic Wavelet Packet Decomposition Based on Frequency-Varying Modulated Lapped Transforms," in *Proc. ICASSP-96*, pp. 1021-1024, May. 1996.
108. D. Sinha and A. Tewfik, "Low Bit Rate Transparent Audio Compression using Adapted Wavelets," *IEEE Trans. Sig. Proc.*, pp. 3463-3479, Dec. 1993.
109. J. Princen and J.D. Johnston, "Audio Coding with Signal Adaptive Filterbanks," in *Proc. ICASSP-95*, pp. 3071-3074, May. 1995.
110. D. Sinha ahd J.D. Johnston, "Audio Compression at Low Bit Rates Using a Signal Adaptive Switched Filterbank," in *Proc. ICASSP-96*, pp. 1053-1056, May. 1996.
111. L. Mainard and M. Lever, "A Bi-Dimensional Coding Scheme Applied to Audio Bitrate Reduction," in *Proc. ICASSP-96*, pp. 1017-1020, May. 1996.
112. S. Boland and M. Deriche, "High Quality Audio Coding Using Multipulse LPC and Wavelet Decomposition," in *Proc. ICASSP-95*, pp. 3067-3070, May. 1995.
113. P. Monta and S. Cheung, "Low Rate Audio Coder with Hierarchical Filterbanks and Lattice Vector Quantization," in *Proc. ICASSP-94*, pp. II-209-II-212, May. 1994.
114. D. Schulz, "Improving Audio Codecs by Noise Substitution," *J. Audio Eng. Soc.*, pp. 593-598, Jul./Aug. 1996.
115. B. Grill, et al., "Improved MPEG-2 Audio Multi-Channel Encoding," in *Proc. 96th Conv. Aud. Eng. Soc.*, preprint #3865, Feb. 1994.
116. W.R. Th. ten Kate, "Scalability in MPEG Audio Compression: From Stereo via 5.1-Channel Surround Sound to 7.1-Channel Augmented Sound Fields," in *Proc. 100th Conv. Aud. Eng. Soc.*, preprint #4196, May 1996.
117. K. Brandenburg and B. Grill, "First Ideas on Scalable Audio Coding," in *Proc. 97th Conv. Aud. Eng. Soc.*, preprint #3924, Nov. 1994.
118. B. Grill and K. Brandenburg, "Two- or Three-Stage Bit-Rate Scalable Audio Coding System," in *Proc. 99th Conv. Aud. Eng. Soc.*, preprint #4132, Oct. 1995.
119. A. Spanias and T. Painter, "Universal Speech and Audio Coding Using a Sinusoidal Signal Model," ASU-TRC Technical Report 97-xxx-001, Jan. 1997.
120. L. Contin, *et al.*, "Tests on MPEG-4 Audio Codec Proposals," *Sig. Proc.: Image Comm. J.*, Oct. 1996.
121. R. Koenen, *et al.*, "MPEG-4: Context and Objectives," *Sig. Proc.: Image Comm. J.*, Oct. 1996.
122. B. Edler, "Current Status of the MPEG-4 Audio Verification Model Development," in *Proc. 101st Conv. Aud. Eng. Soc.*, Preprint #4376, Nov. 1996.