Perceptual Coding of Digital Audio

TED PAINTER, STUDENT MEMBER, IEEE AND ANDREAS SPANIAS, SENIOR MEMBER, IEEE

During the last decade, CD-quality digital audio has essentially replaced analog audio. Emerging digital audio applications for network, wireless, and multimedia computing systems face a series of constraints such 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.

This paper is organized as follows. First, psychoacoustic principles are described, with the MPEG psychoacoustic signal analysis model 1 discussed in some detail. Next, filter bank design issues and algorithms are addressed, with a particular emphasis placed on the modified discrete cosine transform, a perfect reconstruction cosine-modulated filter bank that has become of central importance in perceptual audio coding. Then, we review methodologies that achieve perceptually transparent coding of FM- and CD-quality audio signals, including algorithms that manipulate transform components, subband signal decompositions, sinusoidal signal components, and linear prediction parameters, as well as hybrid algorithms that make use of more than one signal model. These discussions concentrate on architectures and applications of those techniques that utilize psychoacoustic models to exploit efficiently masking characteristics of the human receiver. Several algorithms that have become international and/or commercial standards receive in-depth treatment, including the ISO/IEC MPEG family (−1, −2, −4), the Lucent Technologies PAC/EPAC/MPAC, the Dolby[1] AC-2/AC-3, and the Sony ATRAC/SDDS algorithms. Then, we describe subjective evaluation methodologies in some detail, including the ITU-R BS.1116 recommendation on subjective measurements of small impairments. This paper concludes with a discussion of future research directions.

Keywords*—*AC-2, AC-3, advanced audio coding (AAC), MPEG, ATRAC, audio coding, audio coding standards, audio signal processing, data compression, digital audio radio (DAR), digital broadcast audio (DBA), filter banks, high-definition TV (HDTV), linear predictive coding, lossy compression, modified discrete cosine transform (MDCT), MP3, MPEG, MPEG-1, MPEG-2, MPEG-4, MPEG audio, multimedia signal processing, perceptual audio coding (PAC), perceptual coding, perceptual model, pseudoquadrature mirror filter (PQMF), psychoacoustic model, psychoacoustics, SDDS, signal compression, signal-processing applications, sinusoidal coding, subband coding, transform coding.

Manuscript received November 17, 1999; revised January 24, 2000. This work was supported in part by the NDTC Committee of Intel Corp. under a Grant.

The authors are with the Department of Electrical Engineering, Telecommunications Research Center, Arizona State University, Tempe, AZ 85287-7206 (e-mail: spanias@asu.edu; painter@asu.edu).

Publisher Item Identifier S 0018-9219(00)03054-1.

[1]"Dolby," "Dolby Digital," "AC-2," "AC-3," and "DolbyFAX" are trademarks of Dolby Laboratories. "Sony Dynamic Digital Sound," "SDDS," "ATRAC," and "MiniDisc" are trademarks of Sony Corporation.

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., generating output audio that 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 disc (CD) in the early 1980's [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 either 44.1 or 48 kHz using pulse code modulation (PCM) with a 16-bit sample resolution. This results in uncompressed data rates of 705.6/768 kbits per second (kb/s) for a monaural channel, or 1.41/1.54 Mbits per second (Mb/s) 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 and cost constraints that 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. Therefore, new network and wireless multimedia digital audio systems must reduce data rates without compromising reproduction quality. These and other considerations have motivated considerable research during the last decade toward formulation of compression schemes that can satisfy simultaneously the conflicting demands of high compression ratios and transparent reproduction quality for high-fidelity audio signals [2]–[11]. As a result, several standards have been developed [12]–[15], particularly in the last five years

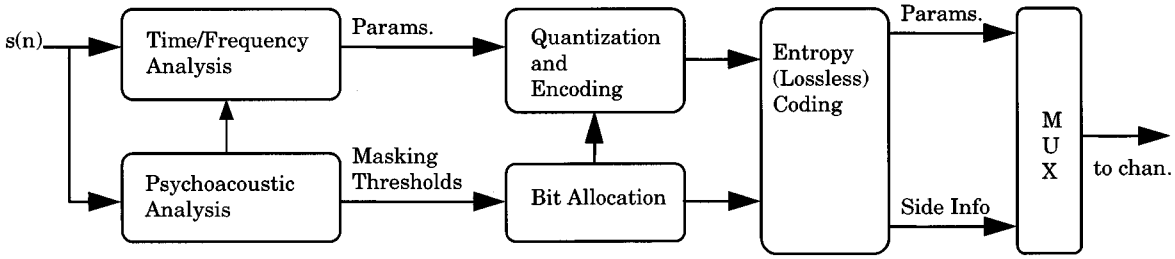


Fig. 1. Generic perceptual audio encoder.

[16]–[19], and several are now being deployed commercially [359], [362], [365], [367] (Table 4).

A. Generic Perceptual Audio Coding Architecture

This review considers several classes of analysis–synthesis data compression algorithms, including those that manipulate transform components, time-domain sequences from critically sampled banks of bandpass filters, sinusoidal signal components, linear predictive coding (LPC) model parameters, or some hybrid parametric set. Within each algorithm class, either lossless or lossy compression is possible. A *lossless* or *noiseless* coding system is able to reconstruct perfectly the samples of the original signal from the coded (compressed) representation. In contrast, a coding scheme incapable of perfect reconstruction from the coded representation is denoted *lossy*. For most audio program material, 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). 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 that seek to exploit the psychoacoustic principles described in Section II. 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/16 bit) 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. The latter debate is beyond the scope of this review.

Before considering different classes of audio coding algorithms, we note the architectural similarities that 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*. Most of these algorithms are based on the generic architecture shown in Fig. 1. The coders typically segment input signals into quasistationary frames ranging from 2 to 50 ms in duration. Then, a time-frequency analysis section estimates the temporal and spectral components on each frame. Often, the time-frequency mapping is matched to the analysis properties of the human auditory system, although this is not always the case. Either way, the ultimate objective is to extract from the input audio a set of time-frequency parameters that is amenable to quantization and encoding in accordance with 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 critically sampled, uniform, or nonuniform bandpass filters;
- time-varying (signal-adaptive) bank of critically sampled, uniform, or nonuniform bandpass filters;
- harmonic/sinusoidal analyzer;
- source-system analysis (LPC/multipulse excitation);
- hybrid transform/filter bank/sinusoidal/LPC signal analyzer.

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 that estimates signal masking power based on psychoacoustic principles (see Section II). The psychoacoustic model delivers masking thresholds that quantify the maximum amount of distortion at each point in the time-frequency plane such that quantization of the time-frequency parameters does not introduce audible artifacts. 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 can be uniform or probability density function (pdf)-optimized (Lloyd–Max), and it might be performed on either scalar or vector data (VQ). Once a quantized compact parametric set has been formed, remaining redundancies are typically removed through noiseless run-length (RL) and entropy (e.g., Huffman [22], arithmetic [23], or Lempel, Ziv, and Welch (LZW) [24], [25]) coding techniques. Since the output of the psychoacoustic distortion control model is signal dependent, most algorithms are inherently variable rate. Fixed channel rate requirements are usually satisfied through buffer feedback schemes, which often introduce encoding delays.

The study of perceptual entropy (PE) suggests that transparent coding is possible in the neighborhood of 2 bits per sample [117] for most for high-fidelity audio sources ($\sim$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 just 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 IV–VIII brings to light many of the tradeoffs involved with the various coder design philosophies.

B. Paper Organization

This paper is organized as follows. In Section II, psychoacoustic principles are described. Johnston's notion of perceptual entropy [45] is presented as a measure of the fundamental limit of transparent compression for audio, and the ISO/IEC MPEG-1 psychoacoustic analysis model 1 is presented. Section III explores filter bank design issues and algorithms, with a particular emphasis placed on the modified discrete cosine transform (MDCT), a perfect reconstruction (PR) cosine-modulated filter bank that is widely used in current perceptual audio coding algorithms. Section III also addresses pre-echo artifacts and control strategies. Sections IV–VII review established and emerging techniques for transparent coding of FM- and CD-quality audio signals, including several algorithms that have become international standards. Transform coding methodologies are described in Section IV, subband coding algorithms are addressed in Section V, sinusoidal algorithms are presented in Section VI, and LPC-based algorithms appear in Section VII. In addition to methods based on uniform bandwidth filter banks, Section V covers coding methods that utilize discrete wavelet transforms (DWT's), discrete wavelet packet transforms (DWPT's), and other nonuniform filter banks. Examples of hybrid algorithms that make use of more than one signal model appear throughout Sections IV–VII. Section VIII is concerned with standardization activities in audio coding. It describes recently adopted standards such as the ISO/IEC MPEG family (−1 ".MP1/2/3," −2, −4), the Phillips' Digital Compact Cassette (DCC), the Sony Minidisk (ATRAC), the cinematic Sony SDDS, the Lucent Technologies Perceptual Audio Coder (PAC)/Enhanced Perceptual Audio Coder (EPAC)/Multichannel PAC (MPAC), and the Dolby AC-2/AC-3. Included in this discussion, Section VIII-A gives complete details on the ".MP3" system, which has been popularized in World Wide Web (WWW) and handheld media applications (e.g., Diamond RIO). Note that the ".MP3" label denotes the MPEG-1, Layer III algorithm. Following the description of the standards, Section IX provides information on subjective quality measures for perceptual codecs. The five-point absolute and differential subjective grading scales are addressed, as well as the subjective test methodologies specified in the ITU-R Recommendation BS.1116. A set of subjective benchmarks is provided for the various standards in both stereophonic and multichannel modes to facilitate interalgorithm comparisons. This paper concludes with a discussion of future research directions.

As an aside, the reader should be aware that the distinction drawn between transform and subband coding in this paper (Sections IV and V) and in the literature is nowadays largely artificial. Although subband versus transform coding class distinctions were justified for the early algorithms that were based on either unitary transforms (e.g., DFT, DCT) or subband filters [e.g., tree-structured quadrature mirror filter (QMF)], the same distinction is not valid for modern algorithms that make use of modulated filter banks such as the MDCT or pseudo-QMF (PQMF). The block transform realizations typically used for the MDCT and PQMF filter banks have been partially responsible for this semantic confusion. A consistent feature of algorithms erroneously lumped into the transform class is that they often make use of very high-resolution filter banks such as a 512-, 1024-, or even 2048-channel MDCT (e.g., ASPEC or DPAC, Sections IV-E and IV-F). Algorithms typically lumped into the subband class tend to make use of lower resolution filter banks, such as a discrete wavelet packet transform with the decomposition tree structured to emulate a critical bandwidth analysis with only 24 subbands (e.g., coders described in Sections V-C and V-D). These consistent (mis)classifications have inspired the logical proposal that the subband/transform class labels for modern coders should be replaced with the classifications of "low-resolution" and "high-resolution" subband coding [33]. The importance of this discussion will become more apparent later in this paper.

For additional information on perceptual coding, one can also refer to informative reviews of recent progress in wideband and high-fidelity audio coding that have appeared in the literature. Discussions of audio signal characteristics and the application of psychoacoustic principles to audio coding can be found in [26]–[28]. Jayant *et al.* of Bell Labs also considered perceptual models and their applications to speech, video, and audio signal compression [29]. Noll describes current algorithms in [30] and [31], including the ISO/MPEG audio compression standards. A recent treatment of the ISO/MPEG algorithms appeared in [75]. Also recently, excellent tutorial perspectives on audio coding fundamentals [32], [62], as well as signal-processing advances [33] central to audio coding, were provided by Brandenburg and Johnston, respectively. In addition, two collections of papers on the current audio coding standards, as well as psychoacoustics, performance measures, and applications, appeared in [34]–[36].

Throughout the remainder of this paper, bit rates will correspond to single-channel or monaural coding, unless otherwise specified. In addition, subjective quality measurements are specified in terms of either the five-point mean opinion score (MOS) or the 41-point subjective difference grade (SDG). These measures are defined in Section IX-A.

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 [37]–[43] has made significant progress toward characterizing human

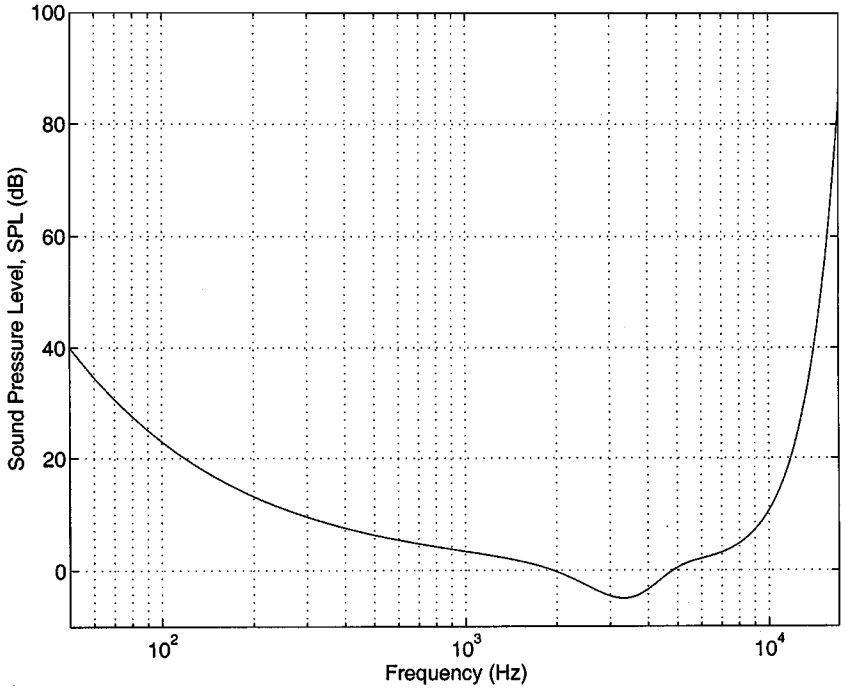


Fig. 2. The absolute threshold of hearing in quiet. Across the audio spectrum, it quantifies the SPL required at each frequency such that an average listener will detect a pure tone stimulus in a noiseless environment.

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 [44], 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 theory of perceptual entropy [45], a quantitative estimate of the fundamental limit of transparent audio signal compression. This section reviews psychoacoustic fundamentals and perceptual entropy, and then gives as an application example some details of the ISO/MPEG psychoacoustic model one.

Before proceeding, however, it is necessary to define the *sound pressure level* (SPL), a standard metric that quantifies the intensity of an acoustical stimulus [42]. Nearly all of the auditory psychophysical phenomena addressed in this paper are treated in terms of SPL. The SPL gives the level (intensity) of sound pressure in decibels (dB) relative to an internationally defined reference level, i.e., $L_{\mathrm{SPL}} = 20 \log_{10}(p/p_0)$ dB, where L_{SPL} is the SPL of a stimulus, p is the sound pressure of the stimulus in Pascals [Pa—equivalent to Newtons per square meter (N/m^2)], and p_0 is the standard reference level of 20 μPa, or 2×10^{-5} N/m^2 [309]. About 150-dB SPL spans the dynamic range of intensity for the human auditory system, from the limits of detection for low-intensity (quiet) stimuli up to the threshold of pain for high-intensity (loud) stimuli. The SPL reference level is calibrated such that the frequency-dependent absolute threshold of hearing in quiet (Section II-A) tends to measure in the vicinity of 0-dB SPL. On the other hand, a stimulus level of 140-dB SPL is typically at or above the threshold of pain. Each of the phenomena addressed in the remainder of this section is characterized in terms of SPL.

A. Absolute Threshold of Hearing

The absolute threshold of hearing characterizes the amount of energy needed in a pure tone such that it can be detected by a listener in a noiseless environment. The absolute threshold is typically expressed in terms of dB SPL. The frequency dependence of this threshold was quantified as early as 1940, when Fletcher [37] reported test results for a range of listeners that were generated in a National Institutes of Health study of typical American hearing acuity. The quiet (absolute) threshold is well approximated [46] by the nonlinear function

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

which is representative of a young listener with acute hearing. When applied to signal compression, $T_q(f)$ could be interpreted naively as a maximum allowable energy level for coding distortions introduced in the frequency domain (Fig. 2). At least two caveats must govern this practice, however. First, whereas the thresholds captured in Fig. 2 are associated with pure tone stimuli, the quantization noise in perceptual coders tends to be spectrally complex rather than tonal. Second, it is important to realize that algorithm

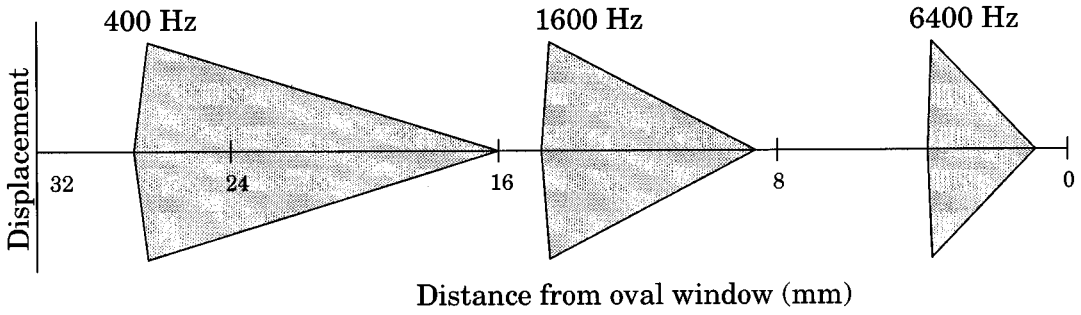


Fig. 3. The frequency-to-place transformation along the basilar membrane. The picture gives a schematic representation of the traveling wave envelopes (measured in terms of vertical membrane displacement) that occur in response to an acoustic tone complex containing sinusoids of 400, 1600, and 6400 Hz. Peak responses for each sinusoid are localized along the membrane surface, with each peak occurring at a particular distance from the oval window (cochlear "input"). Thus, each component of the complex stimulus evokes strong responses only from the neural receptors associated with frequency-specific loci (after [42]).

designers have no a priori knowledge regarding actual playback levels (SPL), and therefore the curve is often referenced to the coding system by equating the lowest point (i.e., near 4 kHz) to the energy in ± 1 bit of signal amplitude. In other words, it is assumed that the playback level (volume control) on a typical decoder will be set such that the smallest possible output signal will be presented close to 0-dB SPL. This assumption is conservative for quiet to moderate listening levels in uncontrolled open-air listening environments, and therefore this referencing practice is commonly found in algorithms that utilize the absolute threshold of hearing. We note that the absolute hearing threshold is related to a commonly encountered acoustical metric other than SPL, namely, dB sensation level (dB SL). *Sensation Level* denotes the intensity level difference for a stimulus relative to a listener's individual unmasked detection threshold for the stimulus [309]. Hence, "equal SL" signal components may have markedly different absolute SPL's, but all equal SL components will have equal suprathreshold margins. The motivation for the use of SL measurements is that SL quantifies listener-specific audibility rather than an absolute level. Whether the target metric is SPL or SL, perceptual coders must eventually reference the internal PCM data to a physical scale. A detailed example of this referencing for SPL is given in Section II-F.

B. Critical Bands

Using the absolute threshold of hearing to shape the coding distortion spectrum represents the first step toward perceptual coding. Considered on its own, however, the absolute threshold is of limited value in the coding context. The detection threshold for spectrally complex quantization noise is a modified version of the absolute threshold, with its shape determined by the stimuli present at any given time. Since stimuli are in general time-varying, the detection threshold is also a time-varying function of the input signal. In order to estimate this threshold, we must first understand how the ear performs spectral analysis. A frequency-to-place transformation takes place in the cochlea (inner ear), along the basilar membrane [42]. The transformation works as follows. A sound wave generated by an acoustic stimulus moves the eardrum and the attached ossicular bones, which in turn transfer the mechanical vibrations to the cochlea, a spiral-shaped, fluid-filled structure that contains the coiled basilar membrane. Once excited by mechanical vibrations at its oval window (the input), the cochlear structure induces traveling waves along the length of the basilar membrane. Neural receptors are connected along the length of the basilar membrane. The traveling waves generate peak responses at frequency-specific membrane positions, and therefore different neural receptors are effectively "tuned" to different frequency bands according to their locations. For sinusoidal stimuli, the traveling wave on the basilar membrane propagates from the oval window until it nears the region with a resonant frequency near that of the stimulus frequency. The wave then slows, and the magnitude increases to a peak. The wave decays rapidly beyond the peak. The location of the peak is referred to as the "best place" or "characteristic place" for the stimulus frequency, and the frequency that best excites a particular place [47], [48] is called the "best frequency" or "characteristic frequency." Thus, a frequency-to-place transformation occurs. An example is given in Fig. 3 for a three-tone stimulus. The interested reader can also find on-line a number of high-quality animations demonstrating this aspect of cochlear mechanics [49]. As a result of the frequency-to-place transformation, the cochlea can be viewed from a signal-processing perspective as a bank of highly overlapping bandpass filters. The magnitude responses are asymmetric and nonlinear (level dependent). Moreover, the cochlear filter passbands are of nonuniform bandwidth, and the bandwidths increase with increasing frequency. The "critical bandwidth" is a function of frequency that quantifies the cochlear filter passbands. Empirical work by several observers led to the modern notion of critical bands [37]–[40]. We will consider two typical examples. In one scenario, the loudness (perceived intensity) remains constant for a narrow-band noise source presented at a constant SPL even as the noise bandwidth is increased up to the critical bandwidth. For any increase beyond the critical bandwidth, the loudness then begins to increase. In this case, one can imagine that loudness remains constant as long as the noise energy is present within only one cochlear

"channel" (critical bandwidth), and then that the loudness increases as soon as the noise energy is forced into adjacent cochlear "channels." Critical bandwidth can also be viewed as the result of auditory detection efficacy in terms of a signal-to-noise ratio (SNR) criterion. The power spectrum model of hearing assumes that masked threshold for a given listener will occur at a constant, listener-specific SNR [50]. In the critical bandwidth measurement experiments, the detection threshold for a narrow-band noise source presented between two masking tones remains constant as long as the frequency separation between the tones remains within a critical bandwidth [Fig. 4(a)]. Beyond this bandwidth, the threshold rapidly decreases [Fig. 4(c)]. From the SNR viewpoint, one can imagine that as long as the masking tones are presented within the passband of the auditory filter (critical bandwidth) that is tuned to the probe noise, the SNR presented to the auditory system remains constant, and hence the detection threshold does not change. As the tones spread further apart and transition into the filter stopband, however, the SNR presented to the auditory system improves, and hence the detection task becomes easier. In order to maintain a constant SNR at threshold for a particular listener, the power spectrum model calls for a reduction in the probe noise commensurate with the reduction in the energy of the masking tones as they transition out of the auditory filter passband. Thus, beyond critical bandwidth, the detection threshold for the probe tones decreases, and the threshold SNR remains constant.

A notched-noise experiment with a similar interpretation can be constructed with masker and maskee roles reversed [Fig. 4(b) and (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. 5(b)] is conveniently approximated [42] by

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

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 that conforms to (2). Table 1 gives an idealized filter bank that corresponds to the discrete points labeled on the curve in Fig. 5(a) and (b). A distance of one critical band is commonly referred to as "one Bark" in the literature. The function [42]

$$z(f) = 13 \arctan(0.00076f) + 3.5 \arctan\left[\left(\frac{f}{7500}\right)^2\right] \text{ (Bark)} \quad (3)$$

is often used to convert from frequency in hertz to the Bark scale [Fig. 5(a)]. Corresponding to the center frequencies of the Table 1 filter bank, the numbered points in Fig. 5(a) illustrate that the nonuniform Hertz spacing of the filter bank (Fig. 6) is actually uniform on a Bark scale. Thus, one critical bandwidth (CB) comprises one Bark.

Although the critical bandwidth captured in (2) is widely used in perceptual models for audio coding, we note that there are alternative expressions. In particular, the equivalent rectangular bandwidth (ERB) scale emerged from research directed toward measurement of auditory filter shapes. In this work, experimental data are obtained typically from notched noise masking procedures. Then, investigators fit the masking data with parametric weighting functions that represent the spectral shaping properties of the auditory filters [50]. Rounded exponential models with one or two free parameters are popular. For example, the single-parameter "roex(p)" model is given by

$$W(g) = (1 + pg)e^{-pg} \quad (4)$$

where

$g = |f - f_0|/f_0$ normalized frequency;
f_0 center frequency of the filter;
f frequency in hertz.

Although the roex(p) model does not capture filter asymmetry, asymmetric filter shapes are possible if two roex(p) models are used independently for the high and low frequency filter skirts. Two parameter models such as the roex(p, r) are also used to gain additional degrees of freedom [50] in order to improve the accuracy of the filter shape estimates. After curve fitting, an ERB estimate is obtained directly from the parametric filter shape. For the roex(p) model, it can be shown easily that the equivalent rectangular bandwidth is given by

$$\text{ERB}_{\text{roex(p)}} = \frac{4f_0}{p}. \quad (5)$$

We note that some texts denote ERB by "equivalent noise bandwidth." An example is given in Fig. 7. The solid line in Fig. 7(a) shows an example roex(p) filter estimated for a center frequency of 1 kHz, while the dashed line shows the ERB associated with the given roex(p) filter shape. In [51] and [52], Moore and Glasberg summarized experimental

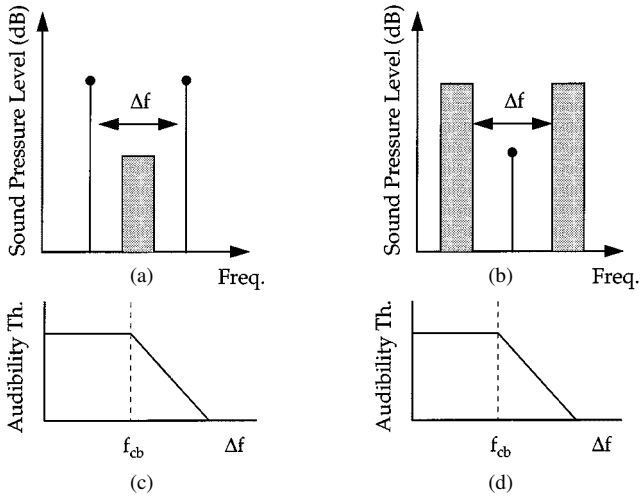


Fig. 4. Critical band measurement methods: (a) and (c) detection threshold decreases as masking tones transition from auditory filter passband into stopband, thus improving detection SNR, and (b) and (d) same interpretation with roles reversed (after [42]).

Table 1
Idealized Critical Band Filter Bank (After [40]). Band Edges and Center Frequencies for a Collection of 25 Critical Bandwidth Auditory Filters That Span the Audio Spectrum. Note That This Idealized Filter Bank Reflects Critical Bandwidth of (2), Not the ERB of (6)

Band No.	Center Freq. (Hz)	Bandwidth (Hz)	Band No.	Center Freq. (Hz)	Bandwidth (Hz)	Band No.	Center Freq. (Hz)	Bandwidth (Hz)
1	50	-100	10	1175	1080-1270	19	4800	4400-5300
2	150	100-200	11	1370	1270-1480	20	5800	5300-6400
3	250	200-300	12	1600	1480-1720	21	7000	6400-7700
4	350	300-400	13	1850	1720-2000	22	8500	7700-9500
5	450	400-510	14	2150	2000-2320	23	10,500	9500-12000
6	570	510-630	15	2500	2320-2700	24	13,500	12000-15500
7	700	630-770	16	2900	2700-3150	25	19,500	15500-
8	840	770-920	17	3400	3150-3700			
9	1000	920-1080	18	4000	3700-4400			

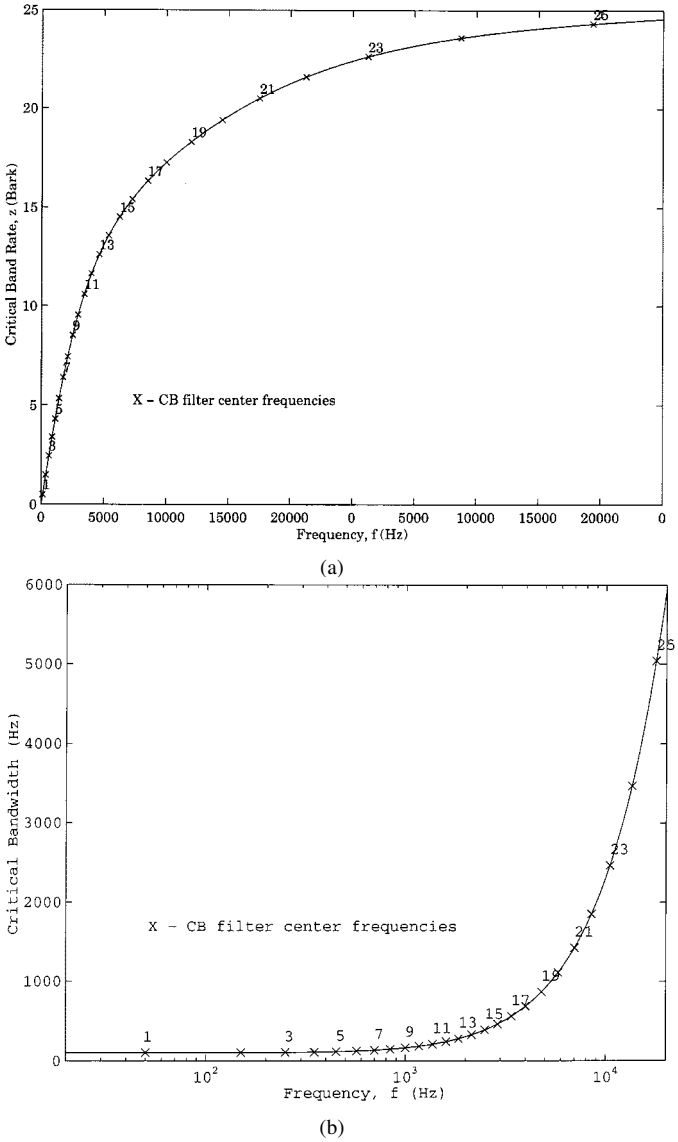


Fig. 5. Two views of critical bandwidth: (a) critical band rate $z(f)$ maps from Hertz to Barks and (b) critical bandwidth $BW_c(f)$ expresses critical bandwidth as a function of center frequency, in Hertz. The X's denote the center frequencies of the idealized critical band filter bank given in Table 1.

ERB measurements for roex(p,r) models obtained over a period of several years by a number of different investigators. Given a collection of ERB measurements on center frequencies across the audio spectrum, a curve fitting on the data set yielded the following expression for ERB as a function of center frequency:

$$\mathrm{ERB}(f) = 24.7(4.37(f/1000) + 1). \tag{6}$$

As shown in Fig. 7(b), the function specified by (6) differs from the critical bandwidth of (2). Of particular interest for perceptual codec designers, the ERB scale implies that auditory filter bandwidths decrease below 500 Hz, whereas the critical bandwidth remains essentially flat. The apparent increased frequency selectivity of the auditory system below 500 Hz has implications for optimal filter bank design, as well as for perceptual bit allocation strategies. These implications are addressed later in this paper.

Regardless or whether it is best characterized in terms of critical bandwidth or ERB, the frequency resolution of the auditory filter bank largely determines which portions of a signal are perceptually irrelevant. The auditory time-frequency analysis that occurs in the critical band filter bank induces simultaneous and nonsimultaneous masking phenomena that are routinely used by modern audio coders to shape the coding distortion spectrum. In particular, the perceptual models allocate bits for signal components such that the quantization noise is shaped to exploit the detection thresholds for a complex sound (e.g., quantization noise). These thresholds are determined by the energy within a critical band [53]. Masking properties and masking thresholds are described next.

C. Simultaneous Masking, Masking Asymmetry, 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 may occur whenever two or more stimuli are simultaneously presented to the auditory system. From a frequency-domain point of view, the relative shapes of the masker and maskee magnitude spectra determine to what extent the presence of certain spectral energy will mask the presence of other spectral energy. From a time-domain perspective, phase relationships between stimuli can also affect masking outcomes. A simplified explanation of the mechanism underlying simultaneous masking phenomena is

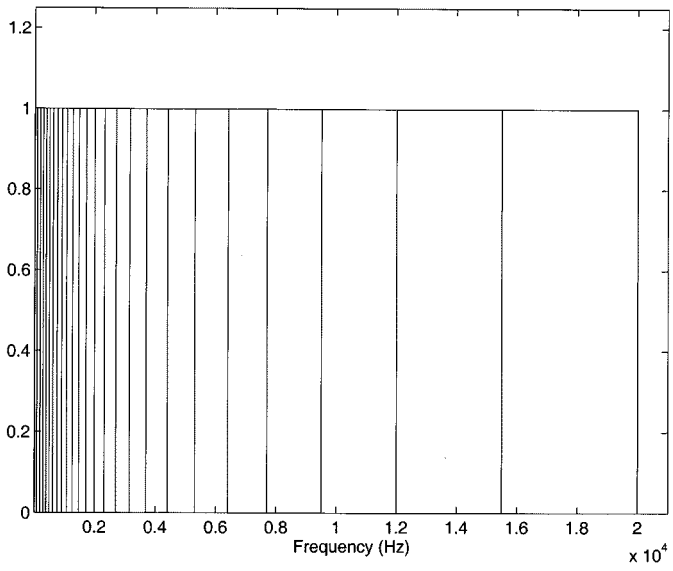


Fig. 6. Idealized critical band filter bank. Illustrates magnitude responses from Table 1. Note that this idealized filter bank reflects critical bandwidth of (2), not the ERB of (6).

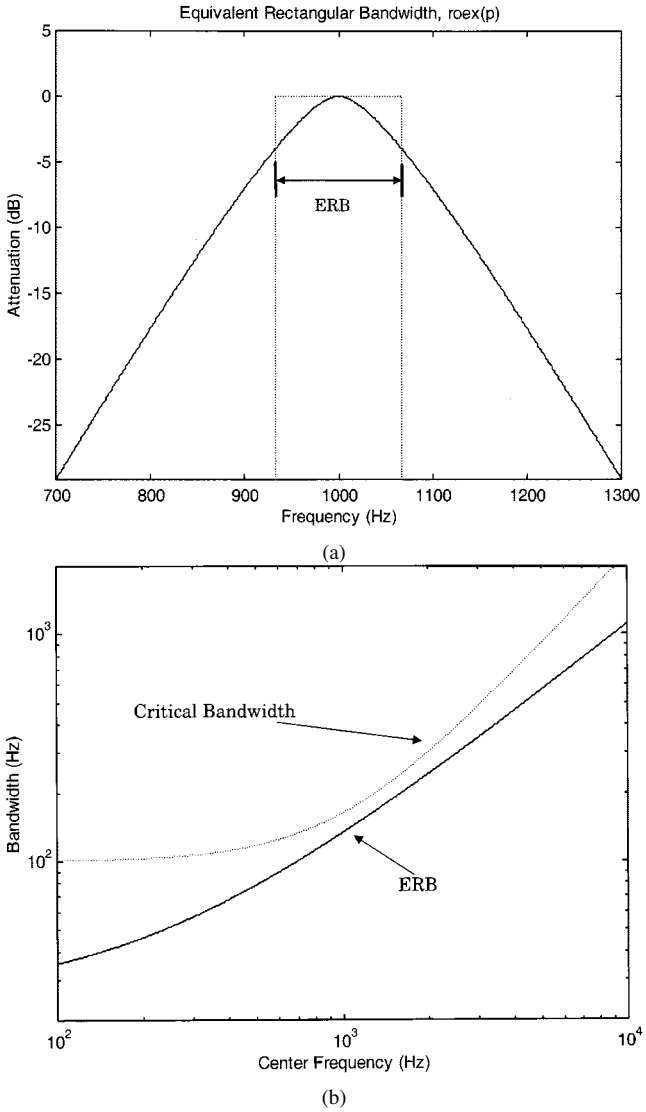


Fig. 7. (a) Example ERB for a roex(p) single-parameter estimate of the shape of the auditory filter centered at 1 kHz. The solid line represents an estimated spectral weighting function for a single-parameter fit to data from a notched noise masking experiment; the dashed line represents the equivalent rectangular bandwidth. (b) ERB versus critical bandwidth—the ERB of (6) (solid) versus critical bandwidth of (2) (dashed) as a function of center frequency.

that 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 block effectively detection of a weaker signal. Although arbitrary audio spectra may contain complex simultaneous masking scenarios, for the purposes of shaping coding distortions it is convenient to distinguish between only three types of simultaneous masking, namely, *noise-masking-tone* (NMT) [40], *tone-masking-noise* (TMN) [41], and *noise-masking-noise* (NMN) [54]. A tutorial treatment of these phenomena and their particular relevance to perceptual coding appeared recently in [54]. Some essential characteristics are described next.

1) Noise-Masking-Tone: In the NMT scenario [Fig. 8(a)], a narrow-band noise (e.g., having 1 Bark bandwidth) masks a tone within the same critical band, provided that the intensity of the masked tone is below a predictable threshold directly related to the intensity—and, to a lesser extent, the center frequency—of the masking noise. Numerous studies characterizing NMT for random noise and pure tone stimuli have appeared since the 1930's (e.g., [55] and [56]). At the threshold of detection for the masked tone, the minimum signal-to-mask ratio (SMR), i.e., the smallest difference between the intensity (SPL) of the masking noise ("signal") and the intensity of the masked tone ("mask") occurs when the frequency of the masked tone is close to the masker's center frequency. In most studies, the minimum SMR tends to lie between -5 and $+5$ dB. For example, a sample threshold SMR result from the NMT investigation [56] is schematically represented in Fig. 8(a). In the figure, a critical band noise masker centered at 410 Hz with an intensity of 80-dB SPL masks a 410-Hz tone, and the resulting SMR at the threshold of detection is 4 dB. Masking power decreases (i.e., SMR increases) for probe tones above and below the frequency of the minimum SMR tone, in accordance with a level- and frequency-dependent spreading function that is described later. We note that temporal factors also affect simultaneous masking. For example, in the NMT scenario, an overshoot effect is possible when the probe tone onset occurs within a short interval immediately following masker onset. Overshoot can boost simultaneous masking (i.e., decrease the threshold minimum SMR) by as much as 10 dB over a brief time span [42]. Section II-D addresses other temporal masking factors.

2) Tone-Masking-Noise: In the case of TMN [Fig. 8(b)], a pure tone occurring at the center of a critical band masks noise of any subcritical bandwidth or shape, provided the noise spectrum is below a predictable threshold directly related to the strength—and, to a lesser extent, the center frequency—of the masking tone. In contrast to NMT, relatively few studies have attempted to characterize TMN. At the threshold of detection for a noise band masked by a

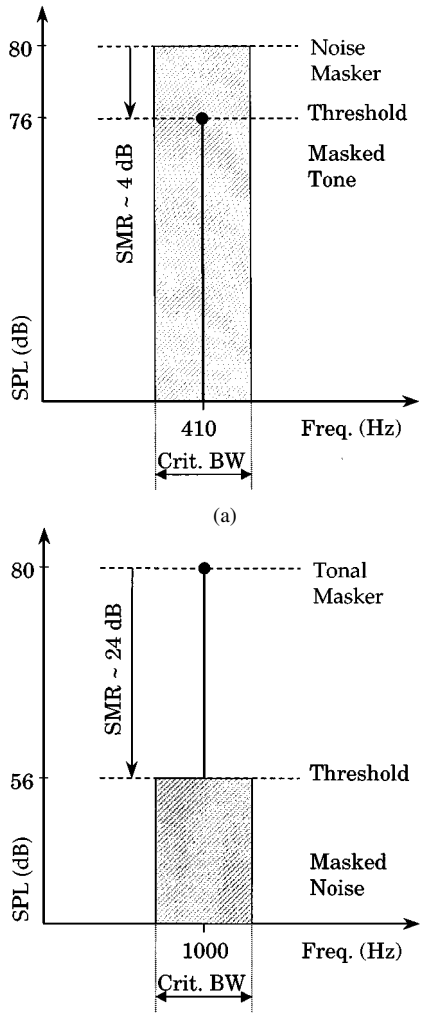


Fig. 8. Example to illustrate the asymmetry of simultaneous masking. (a) Noise-masking-tone—at the threshold of detection, a 410-Hz pure tone presented at 76-dB SPL is just masked by a critical bandwidth narrow-band noise centered at 410 Hz (90-Hz BW) of overall intensity 80-dB SPL. This corresponds to a threshold minimum SMR of 4 dB. The threshold SMR increases as the probe tone is shifted either above or below 410 Hz. (b) Tone-masking-noise—at the threshold of detection, a 1-kHz pure tone presented at 80-dB SPL just masks a critical-band narrow-band noise centered at 1 kHz of overall intensity 56-dB SPL. This corresponds to a threshold minimum SMR of 24 dB. As for the NMT experiment, threshold SMR for the TMN increases as the masking tone is shifted either above or below the noise center frequency 1 kHz. When comparing (a) to (b), it is important to notice the apparent "masking asymmetry," namely, that NMT produces a significantly smaller threshold minimum SMR (4 dB) than does TMN (24 dB). In other words, significantly greater masking power is associated with noise maskers than with tonal maskers. Masking asymmetry is treated in greater depth in [54] and [58].

pure tone, however, it was found in both [41] and [44] that the minimum SMR, i.e., the smallest difference between the intensity of the masking tone ("signal") and the intensity of the masked noise ("mask"), occurs when the masker frequency is close to the center frequency of the probe noise, and that the minimum SMR for TMN tends lie between 21–28 dB. A sample result from the TMN study [44] is given in Fig. 8(b). In the figure, a narrow-band noise of one Bark bandwidth centered at 1 kHz is masked by a 1-kHz tone of intensity 80-dB SPL. The resulting SMR at the threshold of detection is 24 dB. As with NMT, the TMN masking power decreases for critical bandwidth probe noises centered above and below the minimum SMR probe noise.

3) Noise-Masking-Noise: The NMN scenario, in which a narrow-band noise masks another narrow-band noise, is more difficult to characterize than either NMT or TMN because of the confounding influence of phase relationships between the masker and maskee [54]. Essentially, different relative phases between the components of each can lead to different threshold SMR's. The results from one study of intensity difference detection thresholds for wide-band noises [57] produced threshold SMR's of nearly 26 dB for NMN [54].

4) Asymmetry of Masking: The NMT and TMN examples in Fig. 8 clearly show an asymmetry in masking power between the noise masker and the tone masker. In spite of the fact that both maskers are presented at a level of 80-dB SPL, the associated threshold SMR's differ by 20 dB. This asymmetry motivates our interest in both the TMN and NMT masking paradigms, as well as NMN. In fact, knowledge of all three is critical to success in the task of shaping coding distortion such that it is undetectable by the human auditory system. For each temporal analysis interval, a codec's perceptual model should identify across the frequency spectrum noise-like and tone-like components within both the audio signal and the coding distortion. Next, the model should apply the appropriate masking relationships in a frequency-specific manner. In conjunction with the spread of masking (below), NMT, NMN, and TMN properties can then be used to construct a global masking threshold. Although current methods for masking threshold estimation have proven effective, we note that a deeper understanding of masking asymmetry may provide opportunities for improved perceptual models. In particular, Hall [58] has recently shown that masking asymmetry can be explained in terms of relative masker/maskee bandwidths, and not necessarily exclusively in terms of absolute masker properties. Ultimately, this implies that the *de facto* standard energy-based schemes for masking power estimation among perceptual codecs may be valid only so long as the masker bandwidth equals or exceeds maskee (probe) bandwidth. In cases where the probe bandwidth exceeds the masker bandwidth, an envelope-based measure should be embedded in the masking calculation [54], [58].

5) The Spread of Masking: As alluded to earlier, the simultaneous masking effects characterized above by the simplified paradigms of NMT, TMN, and NMN are not band-limited to within the boundaries of a single critical band. Interband masking also occurs, 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 that

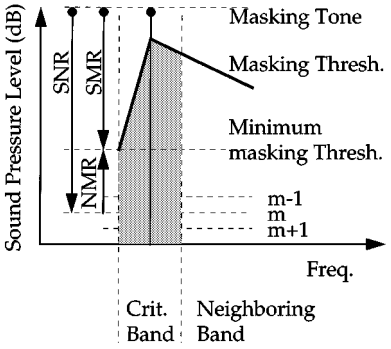


Fig. 9. Schematic representation of simultaneous masking (after [30]).

has slopes of +25 and −10 dB per Bark. A convenient analytical expression [44] is given by

$$SF_{\mathrm{dB}}(x) = 15.81 + 7.5(x + 0.474) - 17.5\sqrt{1 + (x + 0.474)^2}\ \mathrm{dB} \quad (7)$$

where x has units of Barks and $SF_{\mathrm{dB}}(x)$is expressed in dB. After critical band analysis is done and the spread of masking has been accounted for, masking thresholds in perceptual coders are often established by the [59] decibel relations

$$TH_N = E_T - 14.5 - B \quad (8)$$

and

$$TH_T = E_N - K \quad (9)$$

where

TH_N and TH_T	noise and tone masking thresholds, respectively, due to TMN and NMT;
E_N and E_T	critical band noise and tone masker energy levels, respectively;
B	critical band number.

Depending upon the algorithm, the parameter K has typically been set between 3 and 5 dB. Of course, the thresholds of (8) and (9) capture only the contributions of individual tone-like or noise-like maskers. In the actual coding scenario, each frame typically contains a collection of both masker types. One can see easily that (8) and (9) capture the masking asymmetry described previously. After they have been identified, these individual masking thresholds are combined to form a global masking threshold. The global masking threshold comprises an estimate of the level at which quantization noise becomes just noticeable. Consequently, the global masking threshold is sometimes referred to as the level of "just noticeable distortion," or "JND." The standard practice in perceptual coding involves first classifying masking signals as either noise or tone, next computing appropriate thresholds, then using this information to shape the noise spectrum beneath JND. Two illustrated examples are given in Sections II-E and II-F, which are on perceptual entropy, and ISO/IEC MPEG Model 1, respectively. Note that the absolute threshold (T_q) of hearing is also considered when shaping the noise spectra, and that MAX(JND, T_q) 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. 9, 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 that 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. SMR and noise-to-mask ratio (NMR) denote the log distances from the minimum masking threshold to the masker and noise levels, respectively.

D. Nonsimultaneous Masking

As shown in Fig. 10, masking phenomena extend in time beyond the window of simultaneous stimulus presentation. In other words, for a masker of finite duration, nonsimultaneous (also sometimes denoted "temporal") masking occurs both prior to masker onset as well as after masker removal. The skirts on both regions are schematically represented in Fig. 10. Essentially, absolute audibility thresholds for masked sounds are artificially increased prior to, during, and following the occurrence of a masking signal. Whereas significant premasking tends to last only about 1–2 ms, postmasking will extend anywhere from 50 to 300 ms, depending upon the strength and duration of the masker [42]. Tutorial treatments of nonsimultaneous masking have appeared in recent papers on psychoacoustics for audio coding applications [50], [54]. Here we consider key nonsimultaneous masking properties that should be embedded in audio codec perceptual models. Of the two nonsimultaneous masking modes, forward masking is better understood. For masker and probe of the same frequency, experimental studies have shown that the amount of forward (post) masking depends in a predictable way on stimulus frequency [60], masker intensity [60], probe delay after masker cessation [60], and masker duration [50]. Forward masking also exhibits frequency-dependent behavior similar to simultaneous masking that can be observed when the masker and probe frequency relationship is varied [61]. Although backward (pre) masking has also been the subject of many studies, it is less well understood [50]. As shown in Fig. 10, backward masking decays much more rapidly than forward masking. For example, one study at Thomson Consumer Electronics showed that only 2 ms prior to masker onset, the masked threshold was already 25 dB below the threshold of simultaneous masking [62]. We note, however, that the literature lacks consensus over the maximum time persistence of significant backward masking. Despite the inconsistent results across studies, it is nevertheless generally accepted that the amount of measured backward masking depends significantly on the training of the experimental subjects. For the purposes of perceptual coding, abrupt audio signal transients (e.g., the onset of a percussive musical instrument) create pre- and postmasking regions in time

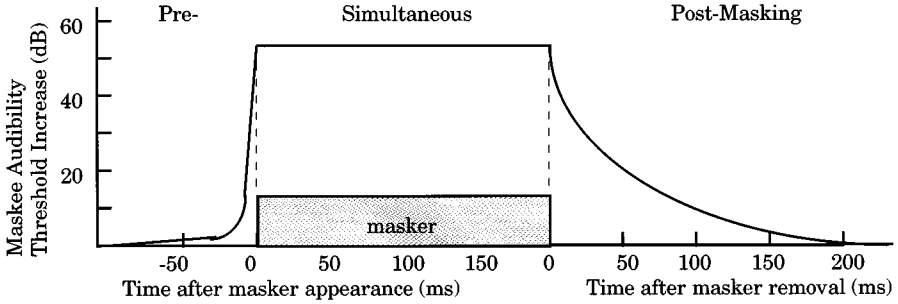


Fig. 10. Nonsimultaneous masking properties of the human ear. Backward (pre) masking occurs prior to masker onset and lasts only a few milliseconds; forward (post) masking may persist for more than 100 ms after masker removal (after [42]).

during which a listener will not perceive signals beneath the elevated audibility thresholds produced by a masker. In fact, temporal masking has been used in several audio coding algorithms (e.g., [12], [63], [112], [268], and [306]). Premasking in particular has been exploited in conjunction with adaptive block size transform coding to compensate for pre-echo distortions (Sections III-D, IV, and VIII).

E. Perceptual Entropy

Johnston, while at Bell Labs, combined notions of psychoacoustic masking with signal quantization principles to define perceptual entropy, 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 [45] 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 Hann window followed by a 2048-point fast Fourier transform (FFT). Masking thresholds are obtained by performing critical band analysis (with spreading), making a determination of the noise-like or tone-like 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

$$P(\omega) = \mathrm{Re}^2(\omega) + \mathrm{Im}^2(\omega) \tag{10}$$

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

$$B_i = \sum_{\omega=bl_i}^{bh_i} P(\omega) \tag{11}$$

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 spreading function (7) is then convolved with the discrete Bark spectrum

$$C_i = B_i * SF_i \tag{12}$$

to account for the spread of masking. An estimation of the tone-like or noise-like quality for C_i is then obtained using the spectral flatness measure (SFM) [64]

$$\mathrm{SFM} = \frac{\mu_g}{\mu_a} \tag{13}$$

where μ_g and μ_a, respectively, correspond to the geometric and arithmetic means of the power spectral density (PSD) components for each band. The SFM has the property that it is bounded by zero and one. Values close to one 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 narrowband. A "coefficient of tonality" α is next derived from the SFM on a dB scale

$$\alpha = \min\left(\frac{SFM_{\mathrm{dB}}}{-60}, 1\right) \tag{14}$$

and this is used to weight the thresholding rules given by (8) and (9) (with $K = 5.5$) as follows for each band to form an offset

$$O_i = \alpha(14.5 + i) + (1 - \alpha)5.5 \text{ (in dB)}. \tag{15}$$

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

$$T_i = 10^{\log_{10}(C_i) - (O_i/10)}. \tag{16}$$

These estimates are scaled by a correction factor to simulate deconvolution of the spreading function, and each T_i is then checked against the absolute threshold of hearing and replaced by $\max(T_i, T_q(i))$. In a manner essentially identical to the SPL calibration procedure that was described in Section II-A, the PE estimation is calibrated by equating the minimum absolute threshold to the energy in a 4-kHz signal of ± 1 bit amplitude. In other words, the system assumes that the playback level (volume control) is configured such that the

smallest possible signal amplitude will be associated with an SPL equal to the minimum absolute threshold. 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

$$\mathrm{PE} = \sum_{i=1}^{25} \sum_{\omega=bl_i}^{bh_i} \log_2 \left(2 \left| n\mathrm{int} \left(\frac{\mathrm{Re}(\omega)}{\sqrt{6T_i/k_i}} \right) \right| + 1 \right) + \log_2 \left(2 \left| n\mathrm{int} \left(\frac{\mathrm{Im}(\omega)}{\sqrt{6T_i/k_i}} \right) \right| + 1 \right) \text{ (bits/sample)} \tag{17}$$

where

i	index of critical band;
bl_i and bh_i	upper and lower bounds of band i;
k_i	number of transform components in band i;
T_i	masking threshold in band i [(16)];
$n\mathrm{int}$	rounding to the nearest integer.

Note that if zero occurs in the log argument, we assign zero 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. In addition, the ISO/IEC MPEG-1 psychoacoustic model 2, which is often used in ".MP3" encoders, is closely related to the PE procedure. We note, however, that there have been evolutionary improvements since the PE estimation scheme first appeared in 1988. For example, the PE calculation in many systems nowadays (e.g., [17]) relies on improved tonality estimates relative to the SFM-based measure of (14). The SFM-based measure is both time and frequency constrained. Only one spectral estimate (analysis frame) is examined in time, and in frequency, the measure by definition lumps together multiple spectral lines. In contrast, the more recently proposed tonality estimation schemes (e.g., the "chaos measure" [17], [62]) consider the predictability of individual frequency components across time, in terms of magnitude and phase tracking properties. A predicted value for each component is compared against its actual value, and the Euclidean distance is mapped to a measure of predictability. Highly predictable spectral components are considered to be tonal, while unpredictable components are treated as noise-like. A tonality coefficient that allows weighting toward one extreme or the other is computed from the chaos measure, just as in (14). Improved performance has been demonstrated in several instances (e.g., [8], [17], [62]). Nevertheless, the PE measurement as proposed in its original form conveys valuable insight on the application of simultaneous masking asymmetry to a perceptual model in a practical system.

F. Example Codec Perceptual Model: 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 I) 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 FFT 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 nontonal 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. We note that although this model is suitable for any of the MPEG-1 coding layers, I-III, the standard [17] recommends that model 1 be used with layers I and II, while model 2 is recommended for layer III (MP3). The five steps leading to computation of global masking thresholds are as follows.

Step 1—Spectral Analysis and SPL Normalization: Spectral analysis and normalization are performed first. The goal of this step is to obtain a high-resolution spectral estimate of the input, with spectral components expressed in terms of sound pressure level. Much like the PE calculation described previously, this SPL normalization guarantees that a 4-kHz signal of ± 1-bit amplitude will be associated with an SPL near 0 dB (close to an acceptable T_q value for normal listeners at 4 kHz), whereas a full-scale sinusoid will be associated with an SPL near 90 dB. The spectral analysis procedure works as follows. 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

$$x(n) = \frac{s(n)}{N(2^{b-1})}. \tag{18}$$

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 PSD estimate $P(k)$ is then obtained using a 512-point FFT, i.e.,

$$P(k) = PN + 10 \log_{10} \left| \sum_{n=0}^{N-1} w(n) x(n) e^{-j(2\pi kn/N)} \right|^2, \quad 0 \le k \le \frac{N}{2} \tag{19}$$

where the power normalization term PN is fixed at 90.302 dB and the Hann window $w(n)$ is defined as

$$w(n) = \frac{1}{2} \left[1 - \cos \left(\frac{2\pi n}{N} \right) \right]. \tag{20}$$

Because playback levels are unknown during psychoacoustic signal analysis, the normalization procedure [(18)] and the parameter PN in (19) are used to estimate SPL conservatively from the input signal. For example, a full-scale

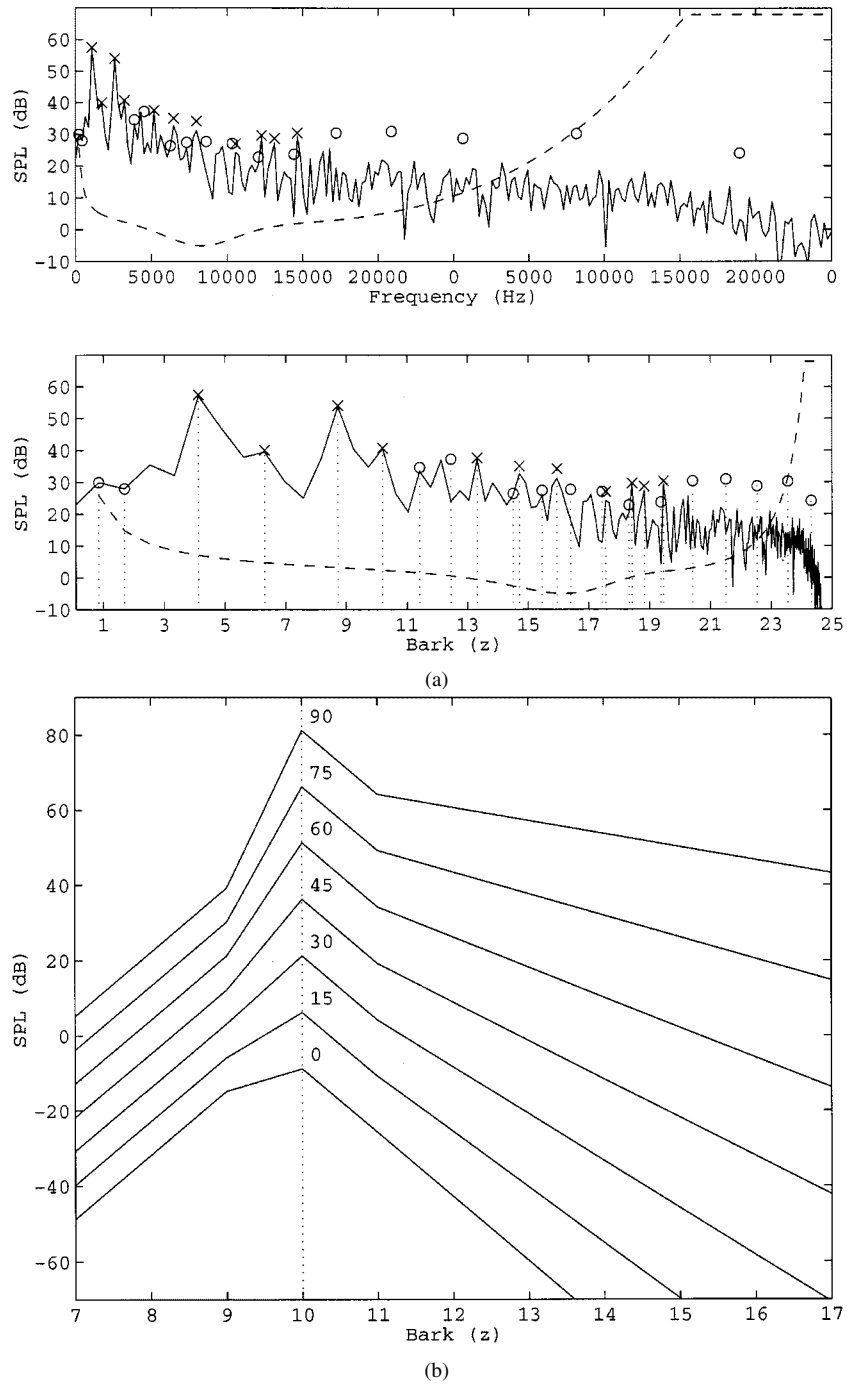
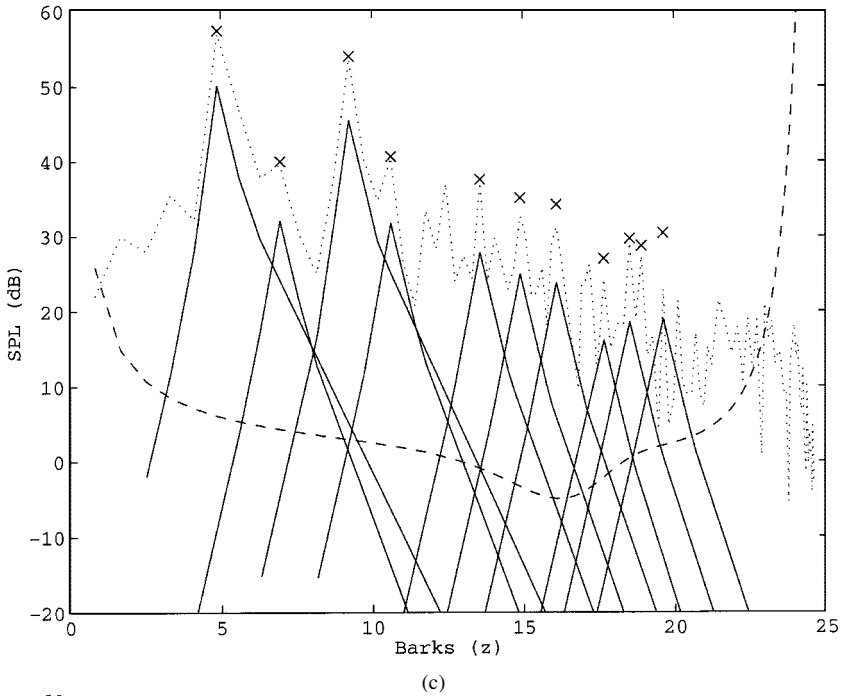


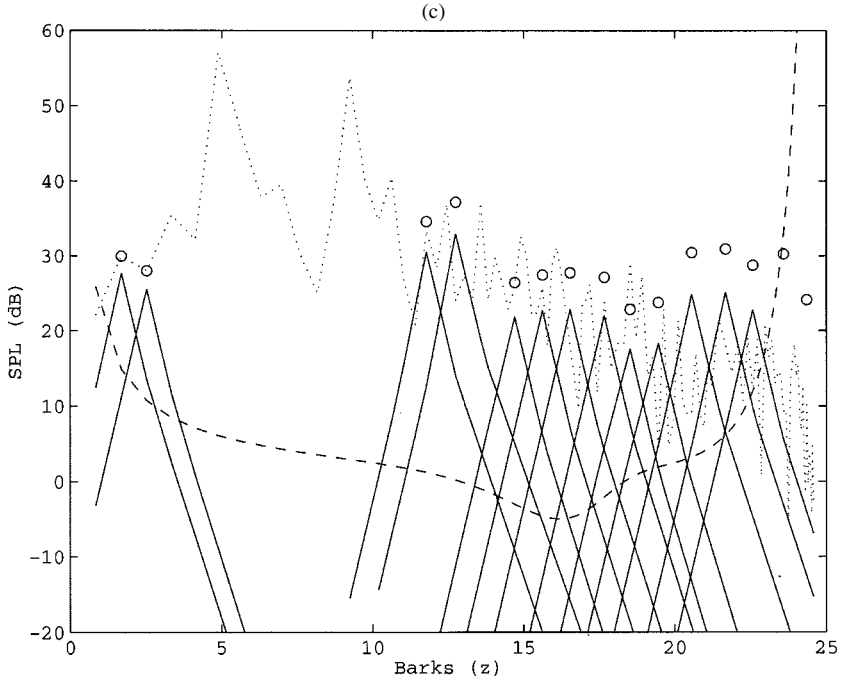
Fig. 11. ISO/IEC MPEG-1 psychoacoustic analysis model 1 for an example pop music selection, steps 1–5 as described in the text. (a) Step 1: Obtain PSD, express in dB SPL. Top panel gives linear frequency scale, bottom panel gives Bark frequency scale. Absolute threshold superimposed. Step 2: Tonal maskers identified and denoted by "X" symbol; Noise maskers identified and denoted by "O" symbol. (b) Collection of prototype spreading functions [(31)] shown with level as the parameter. These illustrate the incorporation of excitation pattern level-dependence into the model. Note that the prototype functions are defined to be piecewise linear on the Bark scale. These will be associated with maskers in steps 3 and 4.

sinusoid that is precisely resolved 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 signals will be at or below the absolute threshold. An example PSD estimate obtained in this manner for a CD-quality pop music selection is given in Fig. 11(a). 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 nontonal masking components are identified. Local maxima in the sample PSD that exceed neighboring components within a



(c)



(d)

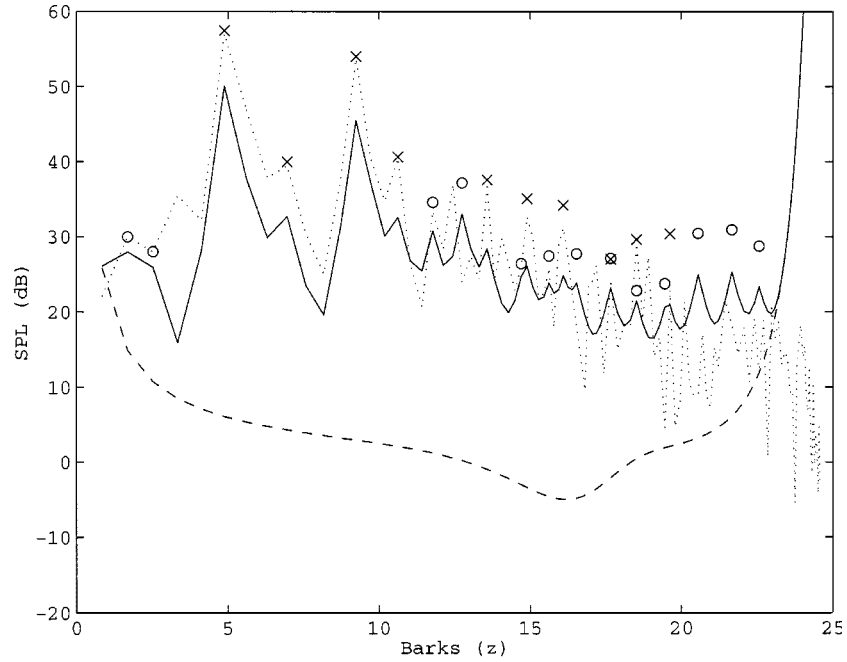
Fig. 11. *(Continued.)* ISO/IEC MPEG-1 psychoacoustic analysis model 1 for an example pop music selection, steps 1–5 as described in the text. (c) Steps 3 and 4: Spreading functions are associated with each of the individual tonal maskers satisfying the rules outlined in the text. Note that the SMR at the peak is close to the widely accepted tonal value of 14.5 dB. (d) Spreading functions are associated with each of the individual noise maskers that were extracted after the tonal maskers had been eliminated from consideration, as described in the text. Note that the peak SMR is close to the widely accepted noise-masker value of 5 dB.

certain Bark distance by at least 7 dB are classified as tonal. Specifically, the "tonal" set S_T is defined as

$$S_T = \left\{ P(k) \,\middle|\, \begin{array}{c} P(k) > P(k \pm 1), \\ P(k) > P(k \pm \Delta_k) + 7 \text{ dB} \end{array} \right\} \qquad (21)$$

where

$$\Delta_k \in \begin{cases} 2 & 2 < k < 63 \quad (0.17\text{–}5.5 \text{ kHz}) \\ [2, 3] & 63 \le k < 127 \quad (5.5\text{–}11 \text{ kHz}) \\ [2, 6] & 127 \le k \le 256 \quad (11\text{–}20 \text{ kHz}) \end{cases} . \qquad (22)$$



(e)

Fig. 11. *(Continued.)* ISO/IEC MPEG-1 psychoacoustic analysis model 1 for an example pop music selection, steps 1–5 as described in the text. (e) Step 5: A global masking threshold is obtained by combining the individual thresholds as described in the text. The maximum of the global threshold and the absolute threshold is taken at each point in frequency to be the final global threshold. The figure clearly shows that some portions of the input spectrum require SNR's of better than 20 dB to prevent audible distortion, while other spectral regions require less than 3-dB SNR. In fact, some high-frequency portions of the signal spectrum are masked and therefore perceptually irrelevant, ultimately requiring no bits for quantization without the introduction of artifacts.

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

$$P_{\mathrm{TM}}(k) = 10\log_{10}\sum_{j=-1}^{1} 10^{0.1P(k+j)}\ (\mathrm{dB}). \tag{23}$$

In other words, for each neighborhood maximum, energy from three adjacent spectral components centered at the peak are combined to form a single tonal masker. Tonal maskers extracted from the example pop music selection are identified using "x" symbols in Fig. 11(a). A single noise masker for each critical band, $P_{\mathrm{NM}}(\overline{k})$, is then computed from (remaining) spectral lines not within the $\pm\Delta_k$ neighborhood of a tonal masker using the sum

$$P_{\mathrm{NM}}(\overline{k}) = 10\log_{10}\sum_{j} 10^{0.1P(j)}\ (\mathrm{dB}),\quad \forall P(j) \notin \{P_{\mathrm{TM}}(k, k\pm 1, k\pm\Delta_k)\} \tag{24}$$

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

$$\overline{k} = \left(\prod_{j=l}^{u} j\right)^{1/(l-u+1)} \tag{25}$$

where l and u are the lower and upper spectral line boundaries of the critical band, respectively. The idea behind (24) is that residual spectral energy within a critical bandwidth not associated with a tonal masker must, by default, be associated with a noise masker. Therefore, in each critical band, (24) combines into a single noise masker all of the energy from spectral components that have not contributed to a tonal masker within the same band. Noise maskers are denoted in Fig. 11 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 that satisfy

$$P_{\mathrm{TM,\,NM}}(k) \geq T_q(k) \tag{26}$$

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. 11(a)] are dropped after application of (26) [Fig. 11(c)–(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 19.5–20.5 Barks [Fig. 11(a)]. It can be seen that the pair is replaced by the stronger of the two during threshold calculations [Fig. 11(c)–(e)]. After the sliding window procedure, masker frequency bins are reorganized according to the subsampling scheme

$$P_{\mathrm{TM,\,NM}}(i) = P_{\mathrm{TM,\,NM}}(k) \tag{27}$$

$$P_{\mathrm{TM,\,NM}}(k) = 0 \tag{28}$$

where

$$i = \begin{cases} k, & 1 \le k \le 48 \\ k + (k \bmod 2), & 49 \le k \le 96 \\ k + 3 - ((k-1) \bmod 4), & 97 \le k \le 232. \end{cases} \tag{29}$$

The net effect of (29) 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 Fig. 11(c)–(e) have been relocated according to this decimation scheme.

Step 4—Calculation of Individual Masking Thresholds: Using the 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 $T_{\mathrm{TM}}(i, j)$ are given by

$$T_{\mathrm{TM}}(i, j) = P_{\mathrm{TM}}(j) - 0.275z(j) + SF(i, j) - 6.025 \text{ (dB SPL)} \tag{30}$$

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

$$SF(i, j) = \begin{cases} 17\Delta_z - 0.4P_{\mathrm{TM}}(j) + 11, & -3 \le \Delta_z < -1 \\ (0.4P_{\mathrm{TM}}(j) + 6)\Delta_z, & -1 \le \Delta_z < 0 \\ -17\Delta_z, & 0 \le \Delta_z < 1 \\ (0.15P_{\mathrm{TM}}(j) - 17)\Delta_z - 0.15P_{\mathrm{TM}}(j), & 1 \le \Delta_z < 8 \end{cases} \text{ (dB SPL)} \tag{31}$$

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 $T_{\mathrm{TM}}(i, j)$ are shown as a function of masker level in Fig. 11(b) for an example tonal masker occurring at $z = 10$ Barks. As shown in Fig. 11, the slope of $T_{\mathrm{TM}}(i, j)$ decreases with increasing masker level. This is a reflection of psychophysical test results, which have demonstrated [42] 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 that occur in the tails of the excitation patterns modeled by $SF(i, j)$. Fig. 11(c) shows the individual masking thresholds [(30)] associated with the tonal maskers in Fig. 11(a) ("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 $T_{\mathrm{NM}}(i, j)$ are given by

$$T_{\mathrm{NM}}(i, j) = P_{\mathrm{NM}}(j) - 0.175z(j) + SF(i, j) - 2.025 \text{ (dB SPL)} \tag{32}$$

where $P_{\mathrm{NM}}(j)$ denotes the SPL of the noise masker in frequency bin j, $z(j)$ denotes the Bark frequency of bin j [(3)], and $SF(i, j)$ is obtained by replacing $P_{\mathrm{TM}}(j)$ with $P_{\mathrm{NM}}(j)$ everywhere in (31). Fig. 11(d) shows the individual masking thresholds associated with the noise maskers identified in step 2 [Fig. 11(a) "o"]. It can be seen in Fig. 11(d) that the two high-frequency noise maskers that occur below the absolute threshold have been eliminated. Before we proceed to step 5 and compute a global masking threshold, it is worthwhile to consider the relationship between (8) and (30), as well as the connection between (9) and (32). Equations (8) and (30) are related in that both model the TMN masking paradigm (Section II-C) in order to generate a masking threshold for quantization noise masked by a tonal signal component. In the case of (8), a Bark-dependent offset that is consistent with experimental TMN data for the threshold minimum SMR is subtracted from the masker intensity, namely, the quantity $14.5 + B$. In a similar manner, (30) estimates for a quantization noise maskee located in bin i the intensity of the masking contribution due the tonal masker located in bin j. Like (8), the psychophysical motivation for (30) is the desire to model the relatively weak masking contributions of a TMN. Unlike (8), however, (30) uses an offset of only $6.025 + 0.275B$, i.e., (30) assumes a smaller minimum SMR at threshold than does (8). The connection between (9) and (32) is analogous. In the case of this equation pair, however, the psychophysical motivation is to model the masking contributions of NMT. Equation (9) assumes a Bark-independent minimum SMR of 3–5 dB, depending on the value of the parameter K. Equation (32), on the other hand, assumes a Bark-dependent threshold minimum SMR of $2.025 + 0.175B$ dB. Also, whereas the spreading function (SF) terms embedded in (30) and (32) explicitly account for the spread of masking, (8) and (9) assume that the spread of masking was captured during the computation of the terms E_T and E_N, respectively.

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 (29). The model assumes that masking effects are additive. The global masking threshold $T_g(i)$ is therefore obtained by computing the sum

$$T_g(i) = 10 \log_{10} \left(10^{0.1T_q(i)} + \sum_{l=1}^{L} 10^{0.1T_{\mathrm{TM}}(i,l)} + \sum_{m=1}^{M} 10^{0.1T_{\mathrm{NM}}(i,m)} \right) \text{ (dB SPL)} \tag{33}$$

where

$T_q(i)$	absolute hearing threshold for frequency bin i;
$T_{\mathrm{TM}}(i, l)$ and $T_{\mathrm{NM}}(i, m)$	individual masking thresholds from step 4;

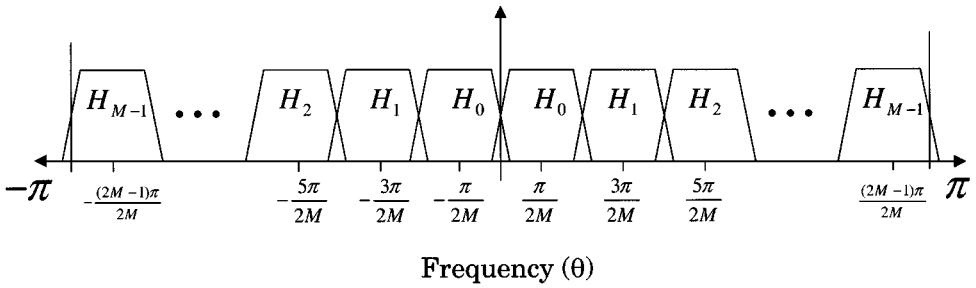


Fig. 12. Magnitude response, oddly stacked uniform M-band filter bank.

L and M numbers 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. Fig. 11(e) shows the global masking threshold obtained by adding the power of the individual tonal [Fig. 11(c)] and noise [Fig. 11(d)] maskers to the absolute threshold in quiet.

III. Time-Frequency Analysis: Filter Banks and Transforms

All audio codecs (Fig. 1) rely upon some type of time-frequency analysis block to extract from the time-domain input a set of parameters that is amenable to quantization and encoding in accordance with a perceptual distortion metric. The tool most commonly employed for this mapping is the filter bank, which is a parallel bank of bandpass filters covering the entire spectrum. The filter bank divides the signal spectrum into frequency subbands and generates a time-indexed series of coefficients representing the frequency-localized signal power within each band. By providing explicit information about the distribution of signal and hence masking power over the time-frequency plane, the filter bank plays an essential role in the identification of perceptual irrelevancies when used in conjunction with a perceptual model. At the same time, the time-frequency parameters generated by the filter bank provide a signal mapping that is conveniently manipulated to shape the coding distortion in order to match the observed time-frequency distribution of masking power. In other words, the filter bank facilitates psychoacoustic analysis as well as perceptual noise shaping. Additionally, by decomposing the signal into its constituent frequency components, the filter bank also assists in the reduction of statistical redundancies. An example magnitude response associated with a uniform bandwidth M-channel filter bank is shown in Fig. 12. The M analysis filters have normalized center frequencies $(2k+1)\pi/2M$, and are characterized by individual impulse responses $h_k(n)$, as well as frequency responses $H_k(\theta)$, for $0 \le k < M$.

Filter banks for audio coding such as the one characterized by the magnitude response of Fig. 12 are perhaps most conveniently described in terms of an analysis–synthesis framework (Fig. 13), in which the input signal $s(n)$ is processed at the encoder by a parallel bank of $(L-1)$th order finite impulse response (FIR) bandpass filters $H_k(z)$. The bandpass analysis outputs

$$\begin{aligned} v_k(n) &= h_k(n) * s(n) \\ &= \sum_{m=0}^{L-1} x(n-m)h_k(m), \quad k = 0, 1, \ldots, M-1 \end{aligned} \tag{34}$$

are decimated by a factor of M, yielding the subband sequences

$$\begin{aligned} y_k(n) &= v_k(Mn) \\ &= \sum_{m=0}^{L-1} x(nM-m)h_k(m), \quad k = 0, 1, \ldots, M-1 \end{aligned} \tag{35}$$

which comprise a *critically sampled* or *maximally decimated* signal representation, i.e., the number of subband samples is equal to the number of input samples. Because it is impossible to achieve perfect "brickwall" magnitude responses with finite order bandpass filters, there is unavoidable aliasing between the decimated subband sequences. Quantization and coding are performed on the subband sequences, $y_k(n)$. In the perceptual audio codec, the quantization noise is usually shaped according to a perceptual model. The quantized subband samples $\hat{y}_k(n)$ are eventually received by the decoder, where they are upsampled by M to form the intermediate sequences

$$w_k(n) = \begin{cases} \hat{y}_k(n/M), & n = 0, M, 2M, 3M, \ldots \\ 0, & \text{otherwise.} \end{cases} \tag{36}$$

In order to eliminate the imaging distortions introduced by the upsampling operations, the sequences $w_k(n)$ are processed by a parallel bank of synthesis filters, $G_k(z)$, and then the filter outputs are combined to form the overall output $\hat{s}(n)$. The analysis and synthesis filters are carefully designed to cancel aliasing and imaging distortions. It can be shown [69] that the overall transfer function of the filter bank is given by

$$\begin{aligned} \hat{s}(n) = \frac{1}{M} \sum_{m=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \sum_{k=0}^{M-1} \\ \cdot s(m)h_k(lM-m)g_k(l-Mn). \end{aligned} \tag{37}$$

For perfect reconstruction filter banks, the output $\hat{s}(n)$ will be identical to the input $s(n)$ within a delay, i.e., $\hat{s}(n) =$

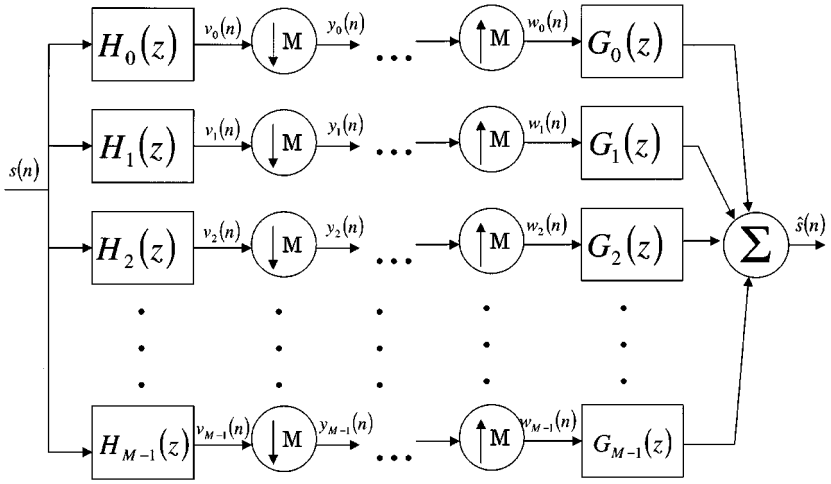


Fig. 13. Uniform M-band maximally decimated analysis–synthesis filter bank.

$s(n - n_0)$, as long as there is no quantization noise introduced, that is, as long as $y_k(n) = \hat{y}_k(n)$. This is naturally not the case for a codec, and therefore quantization sensitivity is an important filter bank property, since PR guarantees are lost in the presence of quantization.

This section provides a perspective on filter bank design considerations, architectures, and special techniques of particular importance in audio coding. This section is organized as follows. First, filter bank design issues for audio coding are addressed. Next, important details on the M-band pseudo-QMF and MDCT filter banks are given. The MDCT is a PR cosine modulated filter bank that has become of central importance in modern audio compression algorithms. Finally, the time-domain "pre-echo" artifact is examined in conjunction with pre-echo control techniques. Beyond the references cited below, the reader in need of greater detail or further analytical development is referred to in-depth tutorials on filter banks that have appeared in the literature [65], [66], as well as in classical [67] and recent texts [68]–[70]. The reader may also wish to explore the connection between filter banks and wavelets that has been well documented in the literature [71], [72] and in several texts [69], [73], [74], [152]. These notions are of particular relevance in the case of audio codecs that make use of discrete wavelet and wavelet packet analysis.

A. Filter Banks for Audio Coding: Design Considerations

The choice of an appropriate filter bank is critical to the success of a perceptual audio coder. Efficient coding performance depends heavily on adequately matching the properties of the analysis filter bank to the characteristics of the input signal [75]. Algorithm designers face an important and difficult tradeoff between time and frequency resolution when selecting a filter bank structure [76]. Failure to choose a suitable filter bank can result in perceptible artifacts in the output (e.g., pre-echoes) or impractically low coding gain and attendant high bit rates. No single resolution tradeoff is optimal for all signals. This dilemma is illustrated in Fig. 14 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. To realize maximum coding gain, the strongly harmonic piccolo signal clearly calls for fine frequency resolution and coarse time resolution, because the masking thresholds are quite localized in frequency, but are

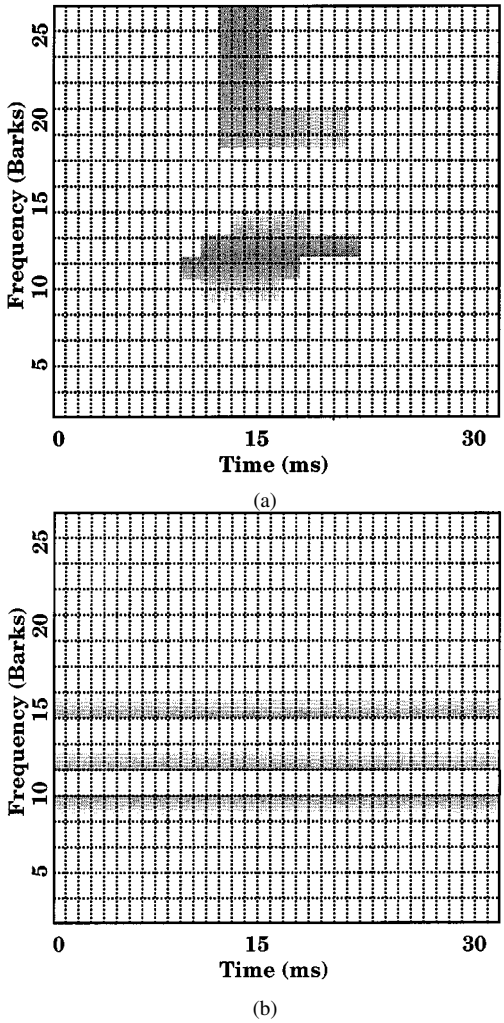


Fig. 14. Masking thresholds in the time-frequency plane: (a) castanets and (b) piccolo (after [201]).

also essentially time-invariant. Quite the opposite is true of the castanets. The fast attacks associated with this percussive sound create highly time-localized masking thresholds that are also widely disbursed in frequency. Therefore, adequate time resolution is essential for accurate estimation of the highly time-varying masked threshold.

Unfortunately, most audio source material is highly nonstationary and contains significant tonal and atonal energy, as well as both steady-state and transient intervals. As a rule, signal models [33] tend to remain constant for long periods and then change suddenly. Therefore, the ideal coder should make adaptive decisions regarding optimal time-frequency signal decomposition, and the ideal analysis filter bank would have time-varying resolutions in both the time and frequency domains. This fact has motivated many algorithm designers to experiment with switched and hybrid filter bank structures, with switching decisions occurring on the basis of the changing signal properties. Filter banks emulating the analysis properties of the human auditory system, i.e., those containing nonuniform "critical bandwidth" subbands, have proven highly effective in the coding of highly transient signals such as the castanets or glockenspiel. For dense harmonically structured signals such as the harpsichord or pitch pipe, on the other hand, the "critical band" filter banks have been less successful because of their reduced coding gain relative to filter banks with a large number of subbands. In short, a number of bank characteristics are highly desirable for audio coding

- signal adaptive time-frequency tiling;
- low-resolution "critical-band" mode, e.g., 32 subbands;
- high-resolution mode, up to 4096 subbands;
- efficient resolution switching;
- minimum blocking artifacts;
- good channel separation;
- strong stopband attenuation;
- perfect reconstruction;
- critical sampling;
- availability of fast algorithms.

Good channel separation and stopband attenuation are particularly desirable for signals containing very little irrelevancy such as the harpsichord. Maximum redundancy removal is essential for maintaining high quality at low bit rates for these signals. Blocking artifacts in time-varying filter banks can lead to audible distortion in the reconstruction. The next two sections, respectively, give some important results on the nearly PR and PR cosine-modulated filter bank architectures that have become of central importance in modern audio coding standards, with particular emphasis on the MDCT. In light of the foregoing discussion on time-frequency resolution, methods for constructing time-varying, signal-adaptive tilings of the time-frequency plane using the MDCT are addressed.

B. Cosine Modulated "Pseudo —QMF" M-Band Banks

Cosine modulation of a lowpass prototype filter has been used since the early 1980's [77]–[81] to realize parallel M-channel filter banks with nearly perfect reconstruction. Because they do not achieve perfect reconstruction, these filter banks are known collectively as "pseudo-QMF," (PQMF) and they are characterized by the following attractive properties:

- constrained design; single FIR prototype filter;
- overall linear phase, and hence constant group delay;
- amenable to fast, block algorithms;
- uniform, linear phase channel responses;
- low complexity = one filter plus modulation;
- critical sampling.

In the PQMF bank derivation [68, ch. 8], phase distortion is completely eliminated from the overall transfer function, (37), because the analysis and synthesis filters are forced to satisfy the mirror image condition

$$g_k(n) = h_k(L-1-n). \tag{38}$$

Moreover, adjacent channel aliasing is cancelled by establishing precise relationships between the analysis and synthesis filters $H_k(z)$ and $G_k(z)$, respectively. In the critically sampled analysis–synthesis notation of Fig. 13, these conditions ultimately yield analysis filters given by

$$h_k(n) = 2w(n)\cos\left[\frac{\pi}{M}(k+0.5)\left(n-\frac{(L-1)}{2}\right)+\theta_k\right] \tag{39}$$

and synthesis filters given by

$$g_k(n) = 2w(n)\cos\left[\frac{\pi}{M}(k+0.5)\left(n-\frac{(L-1)}{2}\right)-\theta_k\right] \tag{40}$$

where

$$\theta_k = (-1)^k\frac{\pi}{4} \tag{41}$$

and the sequence $w(n)$ corresponds to the L-sample "window," a real-coefficient, linear phase FIR prototype low-pass filter, with normalized cutoff frequency $\pi/2M$. Given that aliasing and phase distortions have been eliminated in this formulation, the filter bank design procedure is reduced to the design of the window, $w(n)$, such that overall amplitude distortion is minimized. Examples can be found in [68].

The PQMF bank has played a significant role in the evolution of modern audio codecs. The ISO IS11172-3 and IS13818-3 algorithms ("MPEG-1" [17] and "MPEG-2 BC/LSF" [18]) employ a 32-channel PQMF bank for spectral decomposition in layers I–II. The prototype filter $w(n)$ contains 512 samples, yielding better than 96-dB sidelobe suppression in the stopband of each analysis channel. Output ripple (non-PR) is less than 0.07 dB. In addition, the same PQMF bank is used in conjunction with a PR cosine modulated filter bank in layer III (see Section VI-A) to form a hybrid filter bank architecture with time-varying properties. The MPEG-1 algorithm has reached a position of prominence with the widespread use of ".MP3" files (MPEG-1, layer 3) on the Web for the exchange of audio recordings, as well as with the deployment of MPEG-1, layer II in direct broadcast satellite (DBS/DSS) and European digital audio broadcast (DBA) initiatives. Because of the availability of

common algorithms for PQMF and PR QMF banks, we defer the discussion on generic complexity and efficient implementation strategies until later. In the particular case of MPEG-1, however, note that the 32-band PQMF analysis bank as defined in the standard requires approximately 80 real multiplies and 80 real additions per output sample [17], although a more efficient implementation based on a fast algorithm for the DCT was also proposed [82], [398].

C. Cosine Modulated PR M-Band Banks and the MDCT

Although PQMF banks have been used quite successfully in perceptual audio coders, the overall system design still must compensate for the inherent distortion induced by the lack of perfect reconstruction to avoid audible artifacts in the codec output. The compensation strategy may be a simple one (e.g., increased prototype filter length), but perfect reconstruction is actually preferable because it constrains the sources of output distortion to the quantization stage. Beginning in the early 1990's, independent work by Malvar [83], Ramstad [84], and Koilpillai and Vaidyanathan [85], [86], showed that, in fact, generalized PR cosine modulated filter banks are possible when the prototype low-pass filter $w(n)$ and synthesis filters $g_k(n)$, for $0 \leq k \leq M-1$, are appropriately constrained. These researchers formulated generalized PR cosine modulated filter banks that are of considerable interest in many applications. This section of the paper, however, concentrates on the special case that has become of central importance in the advancement of modern perceptual audio coding algorithms, namely, the filter bank for which $L = 2M$. The PR properties of this special case were first demonstrated by Princen and Bradley [87] using time-domain arguments for the development of the time-domain aliasing cancellation (TDAC) filter bank. Later, Malvar [88] developed the modulated lapped transform (MLT) by restricting attention to a particular prototype filter and formulating the filter bank as a lapped orthogonal block transform. More recently, the consensus name in the audio coding literature for the lapped block transform interpretation of this special-case filter bank has evolved into the modified discrete cosine transform. To avoid confusion, we will denote throughout the remainder of this document by "MDCT" the PR cosine modulated filter bank with $L = 2M$, and we will place some restrictions on the window $w(n)$. In short, the reader should be aware that the different acronyms TDAC, MLT, and MDCT all refer essentially to the same PR cosine modulated filter bank. Only Malvar's MLT label implies a particular choice for $w(n)$, as described below. From the perspective of an analysis–synthesis filter bank (Fig. 13), the MDCT analysis filter impulse responses are given by

$$h_k(n) = w(n)\sqrt{\frac{2}{M}}\cos\left[\frac{(2n+M+1)(2k+1)\pi}{4M}\right] \tag{42}$$

and the synthesis filters, to satisfy the overall linear phase constraint, are obtained by a time reversal, i.e.,

$$g_k(n) = h_k(2M-1-n). \tag{43}$$

This perspective is useful for visualizing individual channel characteristics in terms of their impulse and frequency responses. In practice, however, the filter bank is realized as a block transform.

1) Forward and Inverse MDCT: The analysis filter bank is realized using a block transform of length $2M$ samples and a block advance of only M samples, i.e., with 50% overlap between blocks. Thus, the MDCT basis functions extend across two blocks in time, leading to virtual elimination of the blocking artifacts that plague the reconstruction of nonoverlapped transform coders. Despite the 50% overlap, however, the MDCT is still critically sampled, and only M coefficients are generated by the forward transform for each $2M$-sample input block. Given an input block $x(n)$, the transform coefficients $X(k)$, for $0 \leq k \leq M-1$ are obtained by means of the forward MDCT, defined as

$$X(k) = \sum_{n=0}^{2M-1} x(n)h_k(n). \tag{44}$$

Clearly, the forward MDCT performs a series of inner products between the M analysis filter impulse responses $h_k(n)$ and the input $x(n)$. On the other hand, the inverse MDCT obtains a reconstruction by computing a sum of the basis vectors weighted by the transform coefficients from two blocks. The first M-samples of the kth basis vector, $h_k(n)$, for $0 \leq n \leq M-1$, are weighted by the kth coefficient of the current block, $X(k)$. Simultaneously, the second M-samples of the kth basis vector, $h_k(n)$, for $M \leq n \leq 2M-1$, are weighted by the kth coefficient of the previous block $X^P(k)$. Then, the weighted basis vectors are overlapped and added at each time index n. Note that the extended basis functions require the inverse transform to maintain an M-sample memory to retain the previous set of coefficients. Thus, the reconstructed samples $x(n)$, for $0 \leq n \leq M-1$, are obtained via the inverse MDCT, defined as

$$x(n) = \sum_{k=0}^{M-1} \left[X(k)h_k(n) + X^P(k)h_k(n+M)\right] \tag{45}$$

where $X^P(k)$ denotes the previous block of transform coefficients. The overlapped analysis and overlap-add synthesis processes are illustrated in Fig. 15.

Given the forward [(44)] and inverse [(45)] transform definitions, one still must design a suitable FIR prototype filter $w(n)$. For the MDCT, the generalized PR conditions [68] can be reduced to linear phase and Nyquist constraints on the window, namely

$$w(2M-1-n) = w(n) \tag{46a}$$

and

$$w^2(n) + w^2(n+M) = 1 \tag{46b}$$

for the sample indexes $0 \leq n \leq M-1$. Note that it is possible to modify these constraints and reformulate the MDCT with unique analysis and synthesis windows [89] using a biorthogonal construction. Several general purpose orthogonal [87], [88], [90] and biorthogonal [91]–[93] windows

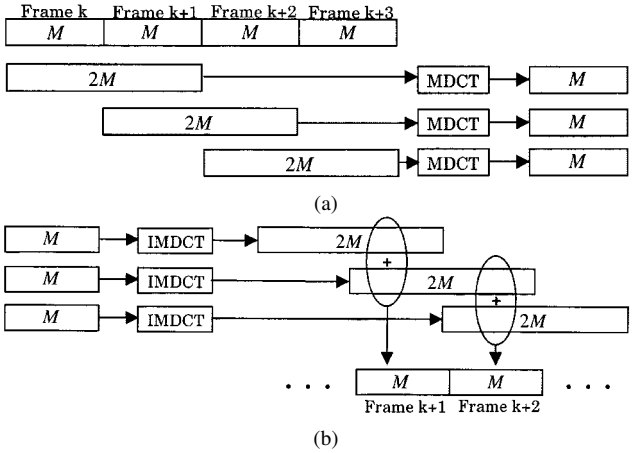


Fig. 15. MDCT: (a) lapped forward transform (analysis)—$2M$ samples are mapped to M spectral components [(44)]. Analysis block length is $2M$ samples, but analysis stride (hop size) and time resolution are M-samples. (b) Inverse transform (synthesis)—M spectral components are mapped to a vector of $2M$ samples [(45)] that is overlapped by M samples and added to the vector of $2M$ samples associated with the previous frame.

have been proposed, while still other orthogonal [94], [112], [268], [362] and biorthogonal [89], [95] windows are optimized explicitly for audio coding.

2) Example Windows: It is instructive to consider some example MDCT windows. Malvar [88] denotes by "MLT" the MDCT filter bank that makes use of the "sine" window, defined as

$$w(n) = \sin\left[\left(n + \frac{1}{2}\right)\frac{\pi}{2M}\right] \tag{47}$$

for $0 \leq n \leq M-1$. This particular window is perhaps the most popular in audio coding. It appears, for example, in the MPEG-1, Layer 3 (MP3) hybrid filter bank [17], the MPEG-2 AAC/MPEG-4 T-F filter bank [112], and numerous experimental coders proposed elsewhere. The sine window has several unique properties that make it advantageous. In particular, dc energy is concentrated in a single coefficient, the filter bank channels have 24-dB sidelobe attenuation, and it can be shown [88] that the MLT is asymptotically optimal in terms of coding gain [64]. Optimization criteria other than coding gain or dc localization have also been investigated. Ferreira [94] proposed a parametric window that offers a controlled tradeoff between reduction of the time-domain ringing artifacts produced by coarse quantization and reduction of stopband leakage relative to the sine window. The Ferreira window has a broader range of better than 110 dB attenuation than does the sine window. Improved ultimate stopband rejection can be beneficial for perceptual gain, particularly for strongly harmonic signals. This realization motivated the designers of the Dolby AC-2/AC-3 [362] and MPEG-2 AAC/MPEG-4 T-F [112] algorithms to use a customized window rather than the standard sine window. The so-called Kaiser–Bessel derived (KBD) window was obtained in a procedure devised at Dolby Laboratories. During the development of the AC-2 and AC-3 algorithms, novel prototype filters were optimized to satisfy a minimum masking template [e.g., Fig. 16(b) for AC-3]. At the expense of some passband selectivity, the KBD windows achieve considerably better stopband attenuation than the sine window [Fig. 16(b)]. Thus, for a pure tone occurring at the center of a particular MDCT channel, the KBD filter bank concentrates more energy into a single transform coefficient. The remaining dispersed energy tends to lie below a worst-case pure tone excitation pattern ["masking template"—Fig. 16(b)]. For signals with adequately spaced tonal components, the presence of fewer suprathreshold MDCT components reduces the perceptual bit allocation.

3) Time-Varying Windows: One final point regarding MDCT window design is of particular importance for perceptual audio coders. As the introduction (Section III-A) illustrated through the pathological cases of tonal and noisy signals, the characteristics of the "best" filter bank for audio are signal specific and therefore time varying. In practice, it is very common for codecs using the MDCT (e.g., MPEG-1 [17], MPEG-2 AAC [112], etc.) to change the window length to match the signal properties of the input. A long window is used to maximize coding gain and achieve good channel separation during segments identified as stationary, and a short window is used to localize time-domain artifacts when pre-echoes are likely. Because of the time overlap between basis vectors, either boundary filters [96] or special transitional windows [97] are required to preserve perfect reconstruction when window switching occurs. Other schemes are also available [98], [99], but for practical reasons these are not typically used. Both the MPEG MDCT-based coders and the Dolby AC-3 algorithm employ MDCT mode switching. Unlike MPEG, however, AC-3 maintains perfect reconstruction without resorting to transitional windows. The spectral and temporal analysis tradeoffs involved in transitional window designs are well illustrated in [106] for both the MPEG-1, layer 3 (MP3) [17] and the Dolby AC-3 [362] filter banks.

4) Fast Algorithms, Complexity, and Implementation Issues: One of the attractive properties that has contributed to the widespread use of the MDCT, particularly in the standards, is the availability of FFT-based fast algorithms [100], [101] that make the filter bank viable for real-time applications. For example, a unified fast algorithm [102] is available for the MPEG-1, -2, -4, and AC-3 long block MDCT, the AC-3 short block MDCT, and the MPEG-1 PQMF bank. A regressive structure suitable for parallel VLSI implementation of the (44) MDCT was also proposed [103]. As far as quantization sensitivity is concerned, there are available expressions [104] for the reconstruction error of the quantized system in terms of signal-correlated and uncorrelated components that can be used to identify perceptually disturbing reconstruction artifacts. Quantization issues for PR cosine modulated filter banks in general are also addressed in [73].

D. Pre-Echo Distortion

An artifact known as pre-echo distortion 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

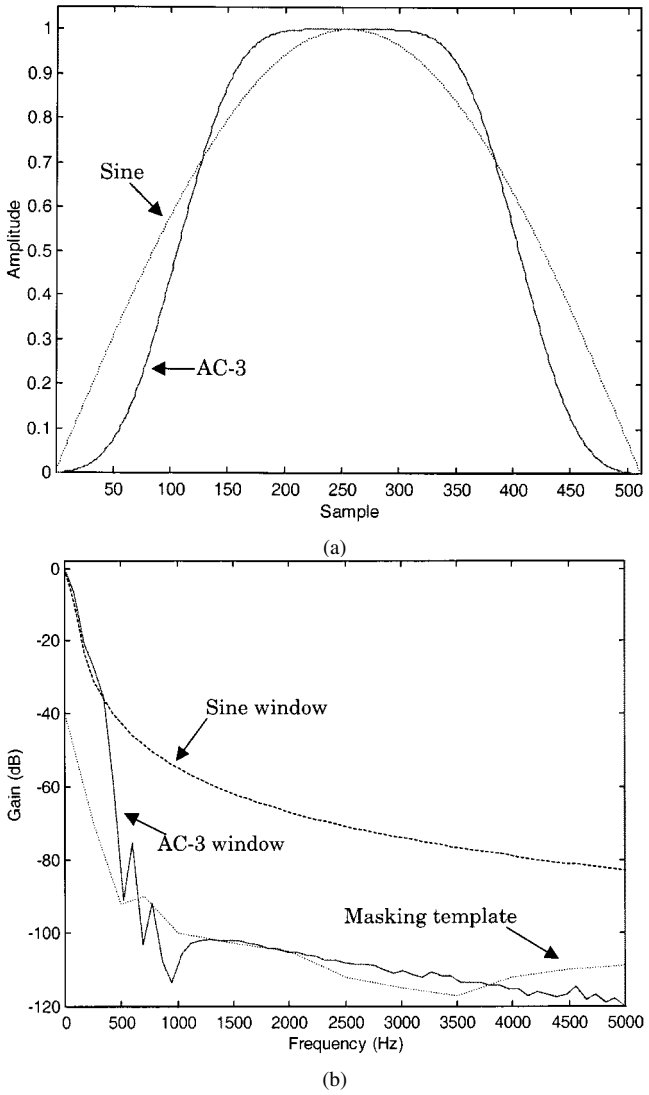


Fig. 16. Dolby AC-3 (solid) versus sine (dashed) MDCT windows: (a) time-domain and (b) magnitude responses in relation to worst case masking template.

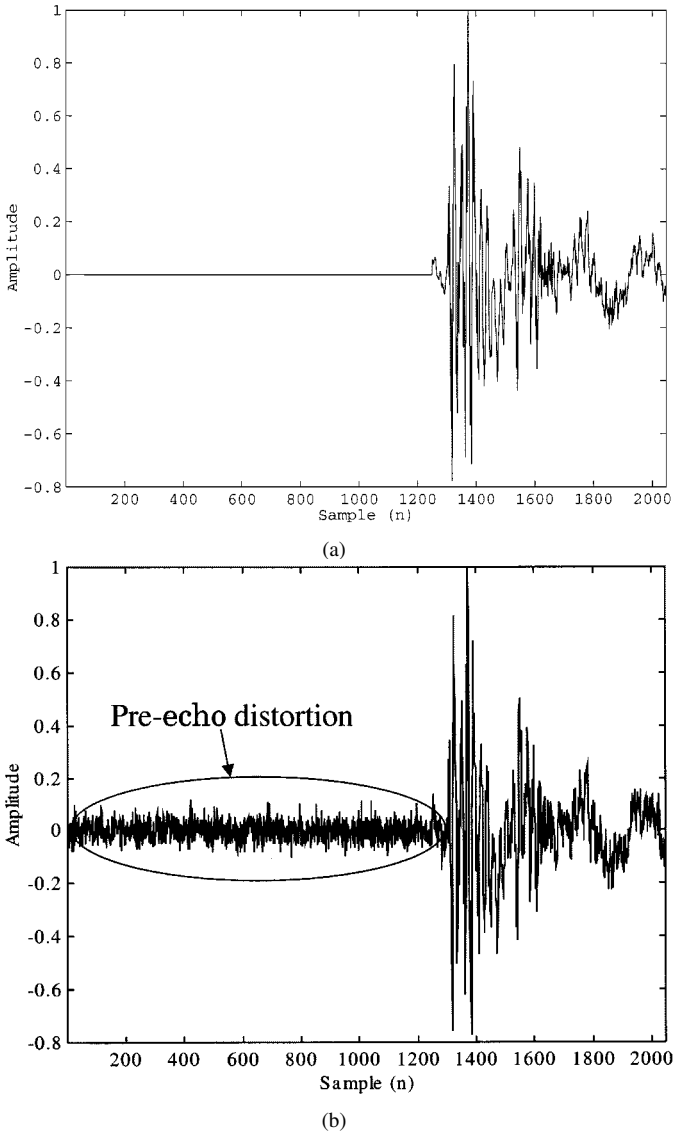


Fig. 17. Pre-echo example: (a) uncoded castanets and (b) transform coded castanets, 2048-point block size.

transform block immediately following a region of low energy. This situation can arise when coding recordings of percussive instruments such as the triangle, the glockenspiel, or the castanets, for example [Fig. 17(a)]. For a block-based algorithm, when quantization and encoding are performed in order to satisfy the masking thresholds associated with the block average spectral estimate, time-frequency uncertainty dictates that the inverse transform will spread quantization distortion evenly in time throughout the reconstructed block [Fig. 17(b)]. This results in unmasked distortion throughout 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 of the distortion is possible only if the transform block size is sufficiently small (minimal coder delay, e.g., 2–5 ms). Percussive sounds are not the only signals likely to produce pre-echoes. Such artifacts also often plague coders when processing "pitched" signals containing nearly impulsive bursts at the beginning of each pitch period, e.g., the "German Male Speech" recording [110]. For a male speaker with a fundamental frequency of 125 Hz, the interval between impulsive events is only 8 ms, which is much less than the typical analysis block length. Several methods proposed to eliminate pre-echoes are reviewed next.

E. Pre-Echo Control Strategies

Several methodologies have been proposed and successfully applied in the effort to mitigate the pre-echoes that tend to plague block-based coding schemes. This section describes several of the most widespread techniques, including the bit reservoir, window switching, gain modification, switched filter banks, and temporal noise shaping. Advantages and drawbacks associated with each method are also discussed.

1) Bit Reservoir: Some coders [17], [307] utilize this technique to satisfy the greater bit demand associated with transients. Although most algorithms are fixed rate, the instantaneous bit rates required to satisfy masked thresholds on each frame are in fact time varying. Thus, the idea behind a bit reservoir is to store surplus bits during periods of

low demand, and then to allocate bits from the reservoir during localized periods of peak demand, resulting in a time-varying instantaneous bit rate but a fixed average bit rate. One problem, however, is that very large reservoirs are needed to deal satisfactorily with certain transient signals, e.g., "pitched signals." Particular bit reservoir implementations are addressed later in conjunction with the MPEG [17] and PAC [307] standards.

2) Window Switching: First introduced by Edler [105], this is also a popular method for pre-echo suppression, particularly in the case of MDCT-based algorithms. Window switching works by changing the analysis block length from "long" duration (e.g., 25 ms) during stationary segments to "short" duration (e.g., 4 ms) when transients are detected. At least two considerations motivate this method. First, a short window applied to the frame containing the transient will tend to minimize the temporal spread of quantization noise such that temporal premasking effects might preclude audibility. Second, it is desirable to constrain the high bit rates associated with transients to the shortest possible temporal regions. Although window switching has been successful [17], [302], [307], it also has significant drawbacks. For one, the perceptual model and lossless coding portions of the coder must support multiple time resolutions. Furthermore, most modern coders use the lapped MDCT. To satisfy PR constraints, window switching typically requires transition windows between the long and short blocks. Even when suitable transition windows (Fig. 18) satisfy the PR constraints, they do so at the expense of poor time and frequency localization properties [106], resulting in reduced coding gain. Other difficulties inherent to window switching schemes are increased coder delay, undesirable latency for closely spaced transients (e.g., long-start–short-stop–start-short), and impractical overuse of short windows for "pitched" signals.

3) Hybrid, Switched Filter Banks: These have also been used to counteract pre-echo distortion. In contrast to window switching schemes, the hybrid and switched filter bank architectures rely upon distinct filter bank modes. In hybrid schemes (e.g., [201]), compatible filter bank elements are cascaded in order to achieve the time-frequency tiling best suited to the current input signal. Switched filter banks (e.g., [308]), on the other hand, make hard switching decisions on each analysis interval in order to select a single monolithic filter bank tailored to the current input. Examples of these methods are given later in this document, along with some discussion of their associated tradeoffs.

4) Gain Modification: This is yet another approach [Fig. 19(a)] that has shown promise in the task of pre-echo control [107], [108]. The gain modification procedure smoothes transient peaks in the time-domain prior to spectral analysis. Then, perceptual coding may proceed as it does for normal, stationary blocks. Quantization noise is shaped to satisfy masking thresholds computed for the equalized long block without compensating for an undesirable temporal spread of quantization noise. A time-varying gain and the modification time interval are transmitted as side information. Inverse operations are performed at the decoder to recover the original signal. Like the other techniques, caveats also apply to this method. For example, gain modification effectively distorts the spectral analysis time window. Depending upon the chosen filter bank, this distortion could have the unintended consequence of broadening the filter bank responses at low frequencies beyond critical bandwidth. One solution for this problem is to apply independent gain modifications selectively within only frequency bands affected by the transient event. This selective approach, however, requires embedding of the gain blocks within a hybrid filter bank structure, which increases coder complexity [109].

5) Temporal Noise Shaping: The final pre-echo control technique considered in this section is temporal noise shaping (TNS). As shown in Fig. 19(b), TNS [110] is a frequency-domain technique that operates on the spectral coefficients $X(k)$ generated by the analysis filter bank. TNS is

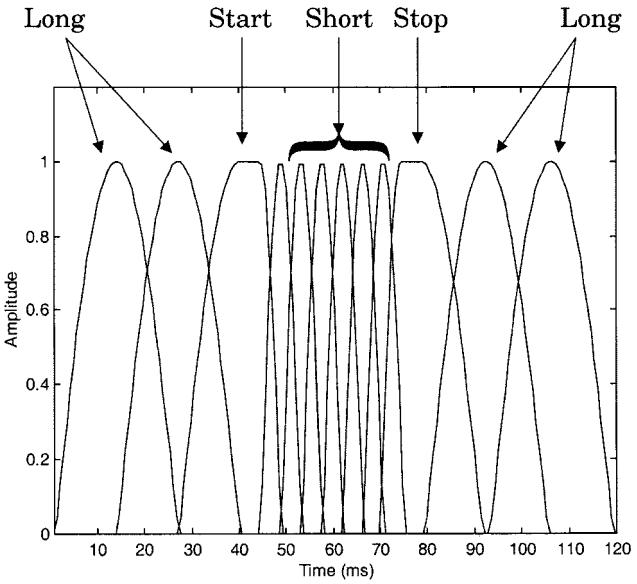


Fig. 18. Example window switching scheme (MPEG-1, Layer III, or "MP3").

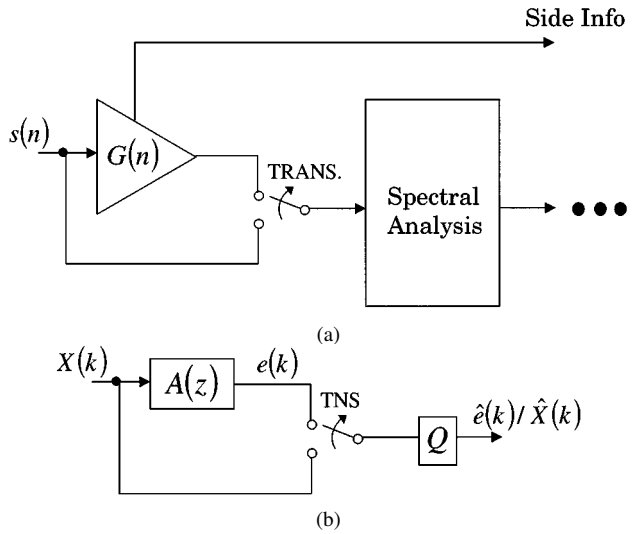


Fig. 19. (a) Gain modification and (b) TNS scheme.

applied only during input attacks susceptible to pre-echoes. The idea is to apply linear prediction (LP) across frequency (rather than time), since for an impulsive time signal, frequency-domain coding gain is maximized using prediction techniques. The method works as follows. Parameters of a spectral LP "synthesis" filter $A(z)$ are estimated via application of standard minimum MSE estimation methods (e.g., Levinson–Durbin [64]) to the spectral coefficients $X(k)$. The resulting prediction residual $e(k)$ is quantized and encoded using standard perceptual coding according to the original masking threshold. Prediction coefficients are transmitted to the receiver as side information to allow recovery of the original signal. The convolution operation associated with spectral domain prediction is associated with multiplication in time. In a manner analogous to the source-system separation realized by LP analysis in the time-domain for traditional speech codecs, therefore, TNS effectively separates the time-domain waveform into an envelope and temporally flat "excitation." Then, because quantization noise is added to the flattened residual, the time-domain multiplicative envelope corresponding to $A(z)$ shapes the quantization noise such that it follows the original signal envelope.

Quantization noise for the castanets applied to a DCT-based coder is shown in Fig. 20(a) and (b) both without and with TNS active, respectively. TNS clearly shapes the quantization noise to follow the input signal's energy envelope. TNS mitigates pre-echoes since the error energy is now concentrated in the time interval associated with the largest masking threshold. Although they are related as time-frequency dual operations, TNS is advantageous relative to gain shaping because it is easily applied selectively in specific frequency subbands. Moreover, TNS has the advantages of compatibility with most filter bank structures and manageable complexity. Unlike window switching schemes, for example, TNS does not require modification of the perceptual model or lossless coding stages to a new time-frequency mapping. TNS was reported in [110] to dramatically improve performance on a five-point mean opinion score (MOS) test from 2.64 to 3.54 for a particularly troublesome pitched signal "German Male Speech" for the MPEG-2 nonbackward compatible (NBC) coder [110]. A MOS improvement of 0.3 was also realized for the well-known "Glockenspiel" test signal. This ultimately led to the adoption of TNS in the MPEG NBC scheme [111], [112].

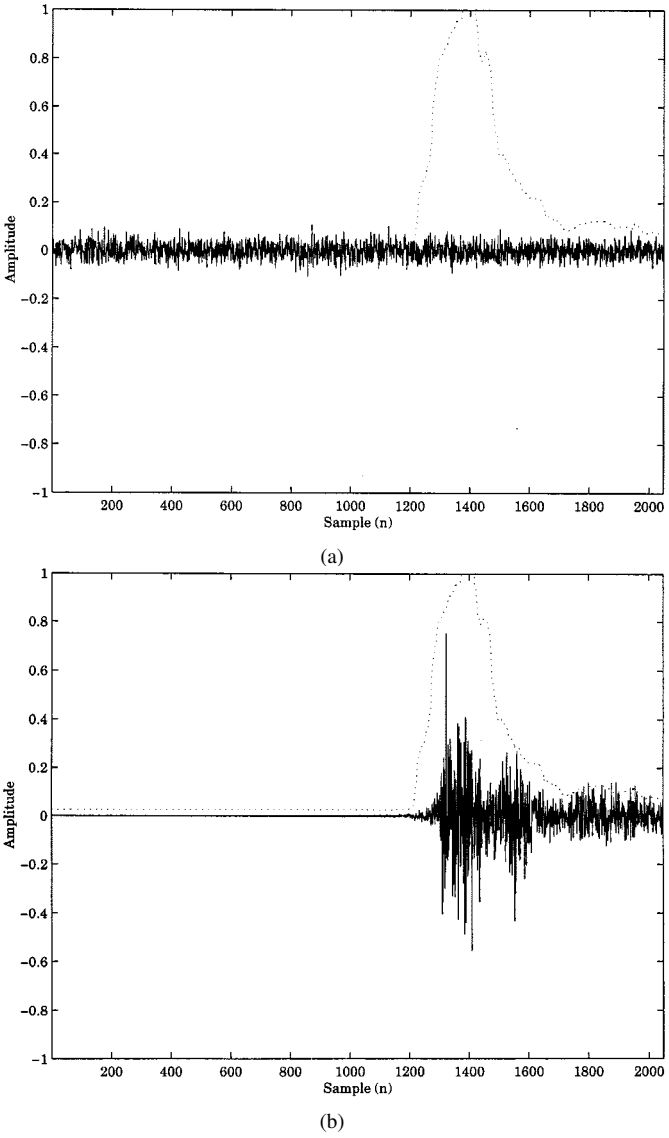


Fig. 20. Temporal noise shaping example showing quantization noise and the input signal energy envelope for castanets: (a) without TNS and (b) with TNS.

IV. 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 schemes for wide-band and high-fidelity audio have been proposed, starting with some of the earliest perceptual audio codecs. In the mid-1980's, Krahe applied psychoacoustic bit allocation principles to a transform coding scheme [113], [114]. Schroeder [3] later extended these ideas into multiple adaptive spectral audio coding (MSC). The MSC utilizes a 1024-point DFT, then groups coefficients into 26 subbands, inspired by the critical bands of the ear. DFT magnitude and phase components are quantized and encoded in a two-step successive refinement procedure that relies upon a perceptual bit allocation. Schroeder reported nearly transparent coding of CD-quality audio at 132 kb/s [3]. Work along these lines has continued, ultimately becoming integral to the current state-of-the-art audio coding standards, although as noted in Section I-B, modern coders making use of the MDCT and other modulated filter banks for high-resolution spectral analysis are in fact subband rather than transform coders. Strictly speaking, the algorithms described in this section that make use of modulated filter banks (e.g., ASPEC, DPAC, TwinVQ) should be called "high-resolution subband coders" rather than transform coders. Also as noted in Section I-B, the source of this confusion has in some

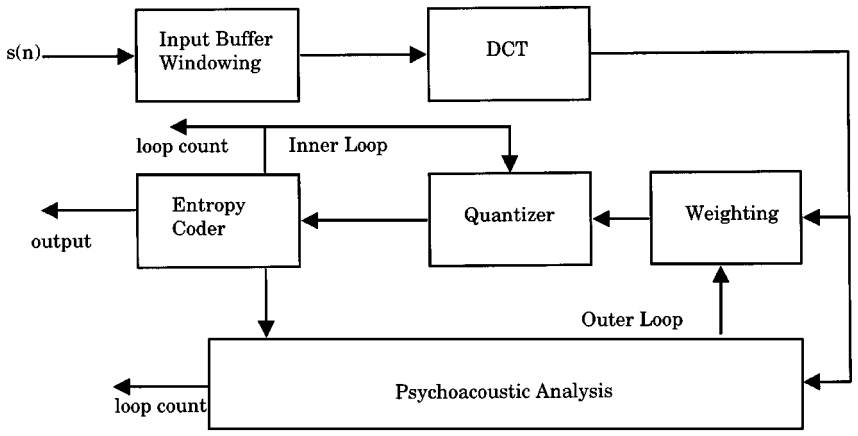


Fig. 21. OCF encoder (after [116]).

cases been the block transform realizations typically used for cosine modulated filter banks. This section describes the individual contributions of Schroeder (MSC) [3], Brandenburg (OCF) [5], [115], [116], Johnston (PXFM/hybrid coder) [6], [8], and Mahieux [118], [119]. Much of this work became connected with MPEG standardization, and ISO/IEC eventually clustered these schemes into a single candidate algorithm, "Adaptive Spectral Entropy Coding of High Quality Music Signals" (ASPEC) [9], which is part of the ISO/IEC MPEG-1 [17] and the MPEG-2/BC-LSF [18] audio coding standards. In fact, most of MPEG-1 Layer III (MP3) and MPEG-2/BC-LSF Layer III is derived from ASPEC. The remainder of this section addresses other novel transform coding schemes that have appeared, not necessarily associated with ASPEC.

A. Optimum Coding in the Frequency Domain (OCF-1, OCF-2, OCF-3)

Brandenburg in 1987 proposed a 132-kb/s 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. 21) works as follows. The input signal is first buffered into 512 sample blocks and transformed to the frequency domain using the DCT. Next, transform components are quantized 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. In the inner loop of Fig. 21, 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.

Using a second iterative procedure, a perceptual analysis is introduced after the inner loop is done. First, critical band analysis is applied. Then, a masking function is applied that combines a flat −6-dB masking threshold with an interband 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 only in the out-of-tolerance critical bands. 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.

Brandenburg in 1988 reported an enhanced OCF (OCF-2), which achieved subjective quality improvements at a reduced bit rate of only 110 kb/s [115]. The improvements were realized by replacing the DCT with the modified DCT (Section III-C) and adding a pre-echo detection/compensation scheme. OCF-2 contains the first reported application of the MDCT to audio coding. The 50% time overlap associated with the MDCT increases the effective time resolution and, consequently, improves the reconstruction quality. OCF-2 quality is also improved for difficult signals such as the triangle and castanets by using a simple pre-echo detection/compensation scheme. OCF-2 was reported to achieve transparency over a wide variety of source material. In 1988, Brandenburg reported further OCF enhancements (OCF-3) in which better quality was realized at a lower bit rate (64 kb/s) with reduced complexity [116]. This was achieved through differential coding of spectral components, an enhanced psychoacoustic model modified to account for temporal masking, and an improved rate-distortion loop.

B. Perceptual Transform Coder (PXFM)

While Brandenburg developed OCF, similar work was simultaneously underway at AT&T Bell Labs. Johnston [6] developed several DFT-based transform coders for audio during the late 1980's that became an integral part of the ASPEC proposal. Johnston's work in perceptual entropy forms the basis for a transform coder reported in 1988 [6] that achieves transparent coding of FM-quality monaural audio signals (Fig. 22). The idea behind the perceptual transform coder (PXFM) is to estimate the amount of quantization noise that can be inaudibly injected into each transform domain subband using PE estimates. The coder 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 I 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

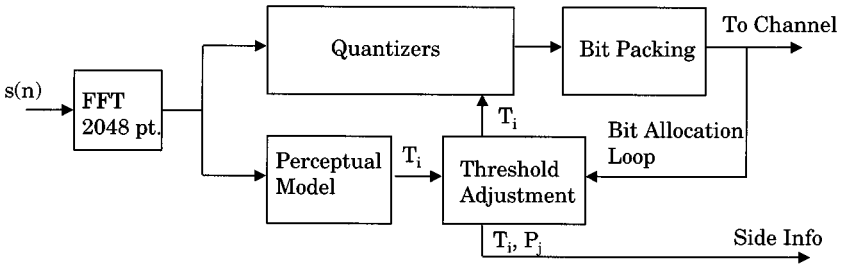


Fig. 22. PXFM encoder (after [6]).

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 eight-sample blocks of complex-valued transform components. 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. SEPXFM [117] realizes performance improvements over PXFM by exploiting inherent stereo cross-channel redundancy. The 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. One of six entropy codebooks is selected for each subband based on the average component magnitude.

C. Brandenburg–Johnston Hybrid Coder

Johnston and Brandenburg [8] collaborated in 1990 to produce a hybrid coder that, strictly speaking, is both a subband and transform coding 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 filter bank that more closely resembled the auditory filter bank. This is accomplished at the encoder by first splitting the input signal into four octave-width subbands using a QMF filter bank. The decimated output sequence from each subband is then followed by one or more transforms to achieve the desired time/frequency resolution [Fig. 23(a)]. Both DFT and MDCT methods were investigated. Given the tiling of the time-frequency plane shown in Fig. 23(b), 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 [(13)] used in PXFM, however, such that tonality is estimated in the hybrid coder as a local characteristic of each individual spectral line. Predictability 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. In contrast, low predictability implies the presence of a noise-like signal. The hybrid coder employs a quantization and coding scheme borrowed from OCF. The hybrid coder without any explicit pre-echo control mechanism was reported to achieve quality better than or equal to OCF-3 at 64 kb/s [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, so too did several CNET researchers. In 1989, Mahieux *et al.* proposed a DFT-based audio coding system that introduced a novel scheme to exploit DFT interblock redundancy. Nearly transparent quality was reported for 15 kHz (FM-grade) audio at 96 kb/s [118], 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 nonuniform pdf-optimized quantizer designed for a Laplacian distribution and an adaptive uniform quantizer, respectively. Bits are allocated during step-size adaptation to shape quantization noise such that a psychoacoustic noise threshold is satisfied for each block. 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. A similar technique was eventually embedded in the MPEG-2 AAC algorithm. In 1990, Mahieux and Petit reported on the development of a similar MDCT-based transform coder for which they reported transparent CD-quality at 64 kb/s [119]. This algorithm introduced a novel "spectrum descriptor" scheme for representing the power spectral envelope. The coder was reported to perform well for broad-band signals with many harmonics but had some problems in the case of spectrally flat signals. More recently, Mahieux and Petit enhanced their 64-kb/s algorithm by incorporating a sophisticated pre-echo detection and postfiltering scheme. Pre-echo postfiltering and improved quantization schemes resulted in a subjective score of 3.65 for two-channel stereo coding at 64 kb/s per channel on the five-point CCIR impairment scale. The CCIR J.41 reference audio codec (MPEG-1, Layer-II) achieved a score of 3.84 at 384 kb/s/channel over the same set of tests.

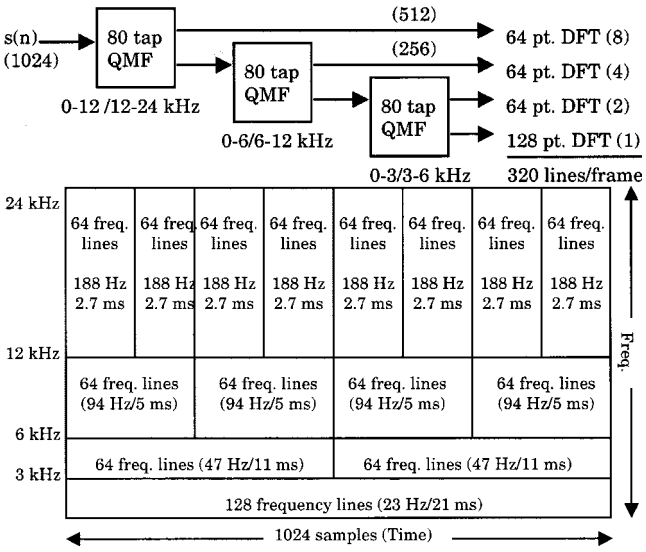


Fig. 23. Brandenburg–Johnston coder: (a) filter bank structure and (b) time/frequency tiling (after [8]).

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 kb/s. 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-echoes are controlled using a dynamic window switching mechanism, like the Thomson coder [105]. ASPEC offers several modes for different quality levels, ranging from 64 to 192 kb/s per channel. ASPEC ultimately formed the basis for Layer III of the MPEG-1 and MPEG-2/BC-LSF standards. We note that similar contributions were made in the area of transform coding for audio outside the ASPEC cluster. For example, Iwadare *et al.* reported on DCT-based [120] and MDCT-based [11] perceptual adaptive transform coders that 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 [121] 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" the frames that contain significant audible differences from the previous frame. The classifier labels nonreference 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 that 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 that utilized three different selection criteria. Best performance was obtained while encoding source material using a PE criterion. As far as overall performance is concerned, NMR measurements were compared between DPAC and Johnston's PXFM algorithm at 64, 88, and 128 kb/s. 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 kb/s, 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, nontransparent applications.

G. DFT Noise Substitution

Other coefficient substitution schemes have also been proposed. Whereas DPAC exploits temporal correlation, a substitution technique that exploits decorrelation was recently devised for coding efficiently noise-like portions of the spectrum. In a noise substitution procedure [122], 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 LP's are applied to the output channels of a multiband QMF analysis filter bank, which has as input the original audio $s(n)$. A predicted signal $\hat{s}(n)$ is obtained by passing the LP output sequences

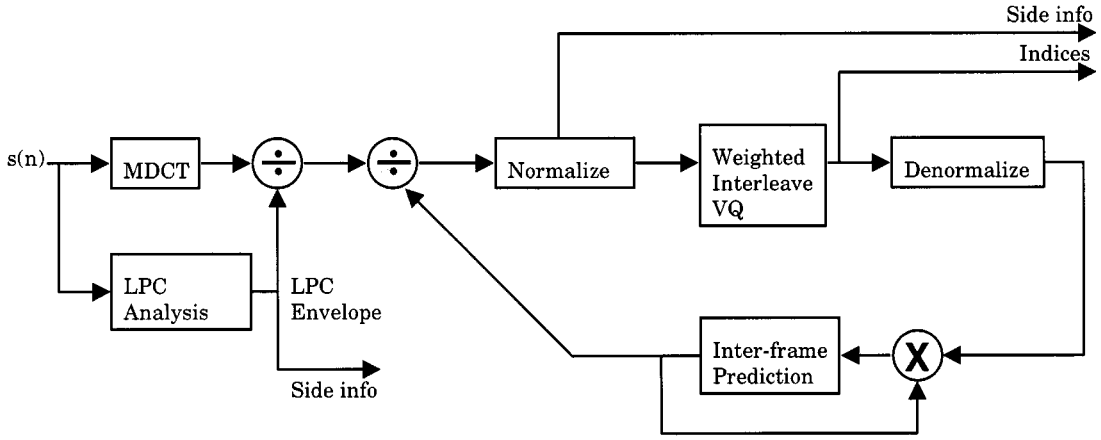


Fig. 24. TWIN-VQ encoder (after [125]).

through the QMF synthesis filter bank. Prediction is done in subbands rather than over the entire spectrum to prevent classification errors that 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.,

$$T(k) = \alpha \left| \frac{S(k) - \hat{S}(k)}{S(k)} \right| + \beta \left| \frac{\theta(k) - \hat{\theta}(k)}{\theta(k)} \right| \qquad (48)$$

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 are offset to a large extent by a significant complexity increase resulting from the additions of the adaptive linear predictors and a multiband analysis–synthesis QMF bank. As a result, the author 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. Gersho and Chan investigated VQ schemes for coding DCT coefficients subject to a constraint of minimum perceptual distortion. They reported on a variable rate coder [7], which achieves high quality in the range of 55–106 kb/s for audio sequences bandlimited to 15 kHz (32 kHz sample rate). After computing the DCT on 512 sample blocks, the algorithm utilizes a novel multistage tree-structured VQ (MSTVQ) scheme for quantization of normalized vectors, with each vector containing four 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 represents a single masker.

Gersho and Chan later enhanced [123] their algorithm by improving the power envelope and transform coefficient quantization schemes. In the new approach to quantization of transform coefficients, constrained-storage VQ (CS-VQ) [124] techniques are combined with the MSTVQ (CS-MSTVQ) from the original coder, allowing the new coder to handle peak NMR requirements without impractical codebook storage requirements. The power envelope samples are encoded using a two-stage process. The first stage applies nonlinear interpolative VQ (NLIVQ). In the second stage, segments of a power envelope residual are encoded using a set of eight-, nine-, and ten-element TSVQ quantizers. Relative to their first VQ/DCT coder, the authors reported savings of 10–20 kb/s 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 [125], [126]. This algorithm exploits LPC analysis, spectral interframe redundancy, and interleaved VQ. At the encoder (Fig. 24), 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. Interleaved VQ is applied to the normalized second step residual. The interleaved 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 ith subvector contains elements $i+nK$, where $n = 0, 1, \ldots, (N/K)-1$. Perceptual weighting is also incorporated by weighting each subvector by a nonlinearly transformed version of its corresponding LPC envelope component prior to the codebook search. VQ indexes are transmitted to the receiver. The authors claimed higher subjective quality than MPEG-1 Layer II at 64 kb/s for 48-kHz CD-quality audio, as well as higher quality than MPEG-1 Layer II for 32-kHz audio at 32 kb/s. More recently, TwinVQ performance at lower bit rates has also been investigated. At least three trends were identified during ISO-sponsored comparative tests [127] of TwinVQ and MPEG-2 AAC (Section VIII-B). First, AAC outperformed TwinVQ for bit rates above 16 kb/s. Second, TwinVQ and AAC achieved similar performance at 16 kb/s, with AAC having a slight edge. Finally, the performance of TwinVQ exceeded that of AAC at a rate of 8 kb/s. These results ultimately motivated a combined AAC/TwinVQ architecture for inclusion in MPEG-4 [385] (Section VIII-C). Enhancements to the weighted interleaving scheme and LPC envelope representation are reported in [128] which enabled real-time implementation of stereo decoders on Pentium and PowerPC platforms. Channel error robustness issues are addressed in [129].

V. Subband Coders

Like the transform coders described in Section IV, 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 either PR or non-PR filter banks (Section III). 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 nonuniform subbands using some critically sampled, PR or non-PR filter bank. Nonideal reconstruction properties in the presence of quantization noise are compensated for by utilizing subband filters that have very good sidelobe attenuation, an approach that usually requires high-order filters. Then, decimated output sequences from the filter bank are normalized and quantized over short, 2–10-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 scale factors 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 filter bank to reconstruct a coded version of the original input.

Numerous subband coding algorithms for high-fidelity audio have appeared in the literature since the late 1980's. In fact, as noted in Section I-B, essentially all modern coders make use of modulated filter banks such as the PQMF or MDCT (Sections III-B and III-C) for high-resolution spectral analysis, particularly for steady-state signals. For analysis of transient signals, on the other hand, a significant number of modern algorithms employ other analysis tools, such as the discrete wavelet packet transform. Typically the DWPT decomposition tree is structured to emulate a (low-resolution) critical band analysis with only 24 subbands (e.g., coders described in Sections V-C and V-D). These trends have inspired the proposal that the subband/transform class labels for modern coders should be replaced with the classifications of "low-resolution" and "high-resolution" subband coding [33]. This section focuses upon the individual subband algorithms proposed by researchers from the Institut fur Rundfunktechnik (IRT) [4], [133], Philips Research Laboratories [134], and CCETT. Much of this work was motivated by standardization activities for the European Eureka-147 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], [135], 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, proposed by Charbonnier and Petit [130], Voros [131], and Teh *et al.* [132], are not discussed here. The first part of this section concentrates upon MUSICAM and its antecedents, which ultimately led to the creation of the MPEG audio standard. The second part of this section describes recent audio coding research in which time-invariant and time-varying signal adaptive filter banks are constructed from DWT's and DWPT's, respectively. This section ends with consideration of some novel hybrid subband/sinusoidal structures that have shown promise.

A. MASCAM

The MUSICAM algorithm is derived from coders developed at IRT, Philips, and CNET. At IRT, Theile *et al.* developed "Masking Pattern Adapted Subband Coding" (MASCAM), a subband audio coder [4] based upon a tree-structured QMF bank that was designed to mimic the critical band structure of the auditory filter bank. The coder has 24 nonuniform subbands, with bandwidths of 125 Hz below 1 kHz, 250 Hz in the range 1–2 kHz, 500

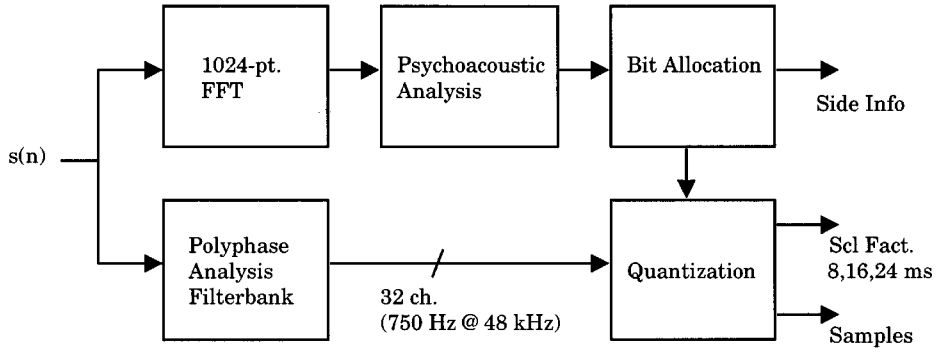


Fig. 25. MUSICAM encoder (after [135]).

Hz in the range 2–4 kHz, 1 kHz in the range 4–8 kHz, and 2 kHz from 8 to 16 kHz. The prototype filter has 64 taps. Subband output sequences are processed in 2-ms blocks. A normalization scalefactor from each subband is quantized and transmitted for each block. 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 [133], interband 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 80–100 kb/s per channel. A similar subband coder was developed at Philips during this same period. As described by Veldhuis *et al.* in [134], the Philips group investigated subband schemes based on 20- and 26-band nonuniform filter banks. Like the original MASCAM system, the Philips coder relies upon a highly simplified masking model that 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 energy 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 kb/s for the 26-band version and 180 kb/s for the 20-band version.

B. MUSICAM

Based primarily upon coders developed at IRT and Phillips, the MUSICAM algorithm [10], [135] was successful in the 1990 ISO/IEC competition [136] 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. 25) makes several practical tradeoffs between complexity, delay, and quality. By utilizing a uniform bandwidth, 32-band polyphase filter bank instead of a tree-structured QMF bank, 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 suboptimal filter bank, however, in the sense that filter bandwidths (constant 750 Hz for 48-kHz sample rate) no longer correspond to the critical bands. 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 filter bank. This parallel structure allows for improved estimation of masking thresholds and hence determination of more accurate minimum SMR's required within each subband. The MUSICAM psychoacoustic analysis procedure is essentially the same as the MPEG-1 psychoacoustic model 1 described in Section VIII-G.

The remainder of MUSICAM works as follows. Subband output sequences are processed in 8-ms blocks (twelve samples at 48 kHz), which is close to the temporal resolution of the auditory system (4–6 ms). Scale factors are extracted from each block and encoded using 6 bits over a 120-dB dynamic range. Occasionally, temporal redundancy is exploited by repetition over two or three blocks (16 or 24 ms) of slowly changing scale factors 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 (standard deviation 0.7) at 128 kb/s, and 4.3 (standard deviation 1.1) at 96 kb/s per monaural channel, compared to 4.7 (standard deviation 0.6) on the same scale for the uncoded original. Quality was reported to suffer somewhat at 96 kb/s 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}.

C. Wavelet Decompositions

The previous section described subband coding algorithms that utilize banks of fixed resolution bandpass QMF or polyphase FIR filters. This section describes a different

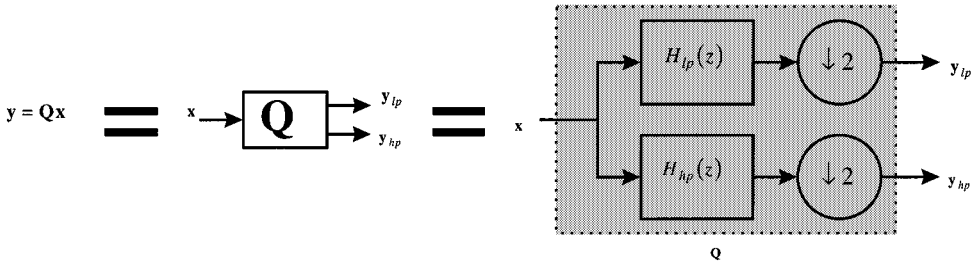


Fig. 26. Filter bank interpretation of the DWT.

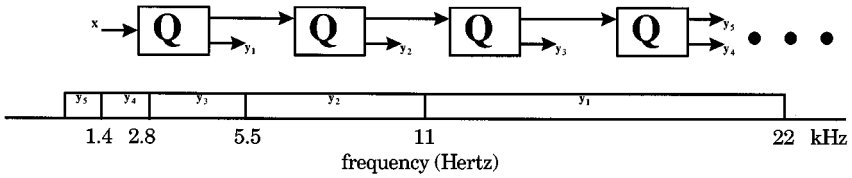


Fig. 27. Subband decomposition associated with a discrete wavelet transform.

class of subband coders that rely instead upon a filter bank interpretation of the DWT. DWT based subband coders offer increased flexibility over the subband coders described previously since identical filter bank magnitude frequency responses can be obtained for many different choices of a wavelet basis, or equivalently, choices of filter coefficients. This flexibility presents an opportunity for basis optimization. For each segment of audio, one can adaptively choose a wavelet basis that minimizes the rate for some target distortion. 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 treatment of wavelets is available from many sources, for example, [137]. 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 $\boldsymbol{Q}$ is fully determined by a set of coefficients $\{c_k\}$ for $0 \leq k \leq K-1$. As shown in Fig. 26, this transformation matrix has an associated filter bank interpretation. One application of the transform matrix $\boldsymbol{Q}$ to an $N\times 1$ signal vector $\boldsymbol{x}$ generates an $N\times$ 1 vector of wavelet-domain transform coefficients $\boldsymbol{y}$. The $N\times 1$ vector $\boldsymbol{y}$ can be separated into two $(N/2)\times 1$ vectors of approximation and detail coefficients $\boldsymbol{y}_{\mathrm{lp}}$ and $\boldsymbol{y}_{\mathrm{hp}}$, respectively. The spectral content of the signal $\boldsymbol{x}$ captured in $\boldsymbol{y}_{\mathrm{lp}}$ and $\boldsymbol{y}_{\mathrm{hp}}$ corresponds to the frequency subbands realized in 2 : 1 decimated output sequences from a QMF bank.

Therefore, recursive DWT applications effectively pass input data through a tree-structured cascade of low-pass and high-pass filters followed by 2 : 1 decimation at every node. The forward/inverse transform matrices of a particular wavelet are associated with a corresponding QMF analysis/synthesis filter bank. The usual wavelet decomposition implements an octave-band filter bank structure shown in Fig. 27. In the figure, frequency subbands associated with the coefficients from each stage are schematically represented for an audio signal sampled at 44.1 kHz.

Wavelet packet (WP) or DWPT representations, on the other hand, decompose both the detail and approximation coefficients at each stage of the tree, as shown in Fig. 28. In the figure, frequency subbands associated with the coefficients from each stage are schematically represented for a 44.1-kHz sample rate. A filter bank interpretation of wavelet transforms is attractive in the context of audio coding algorithms. 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 filter bank utilizing a wavelet packet approach. Moreover, many K-coefficient finite 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 that is in some sense "optimal." The basic idea behind DWT and DWPT-based subband coders is to quantize and encode efficiently the coefficient sequences associated with each stage of the wavelet decomposition tree using the same noise shaping techniques as the previously described perceptual subband coders. The next few subsections concentrate upon WP-based subband coders developed in the early 1990's by Sinha *et al.* [157], [158], [160], as well as more recently proposed hybrid sinusoidal/WPT algorithms developed by Hamdy and Tewfik [187], Boland and Deriche [138], and Pena *et al.* [139]–[142]. 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 [143], Kudumakis and Sandler [144]–[146], and Boland and Deriche [147], [148]. Several bit-rate scalable DWPT-based schemes have also been investigated recently. For example, a fixed-tree DWPT coding scheme capable of nearly transparent quality with scalable bit rates below 100 kb/s was proposed by Dobson *et al.* and implemented in real time on a 75-MHz Pentium-class platform [149]. Additionally, Lu and

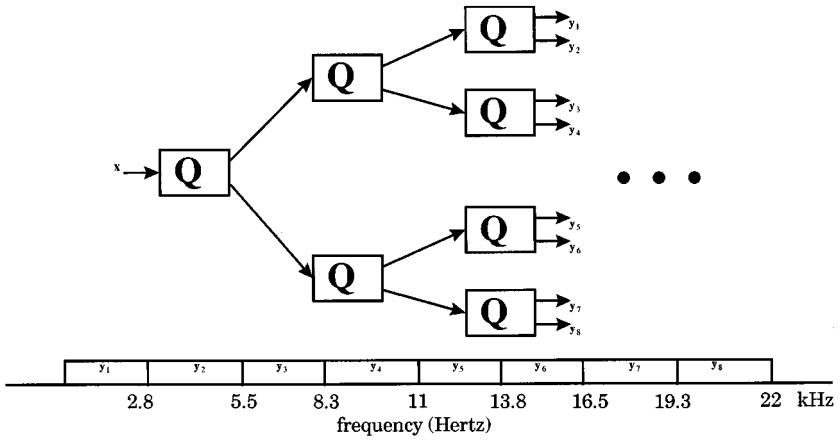


Fig. 28. Subband decomposition associated with discrete wavelet packet transform (DWPT or WP). Note that other, nonuniform decomposition trees are also possible.

Pearlman investigated a rate-scalable DWPT-based coder that applies set partitioning in hierarchical trees (SPIHT) to generate an embedded bitstream. Nearly transparent quality was reported at bit rates between 55–66 kb/s [150].

D. Adapted Wavelet Packet Decompositions

The "best basis" methodologies [151], [152] for adapting the WP *tree structure* to signal properties are typically formulated in terms of Shannon entropy [153] and other perceptually blind statistical measures. For a given WP tree, related research directed toward *optimal filter selection* [154]–[156] has also emphasized optimization of statistical rather than perceptual properties. The questions of perceptually motivated filter selection and tree construction are central to successful application of WP analysis in audio coding algorithms. We consider in this section some relevant research and algorithm developments. The WP tree structure determines the time and frequency resolution of the transform and therefore also creates a particular tiling of the time-frequency plane. Several WP audio algorithms [149], [158] have successfully employed time-invariant WP tree structures that mimic the ear's critical band frequency resolution properties. In some cases, however, a more efficient perceptual bit allocation is possible with a signal-specific time-frequency tiling that tracks the shape of the time-varying masking threshold. Some examples are described next.

1) DWPT Coder with Globally Adapted Daubechies Analysis Wavelet: 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 kb/s [157], [158]. The encoder (Fig. 29) exploits redundancy using a VQ scheme and irrelevancy using a WP signal decomposition combined with perceptual masking thresholds. The algorithm works as follows. Input audio is segmented into $N\times 1$ vectors, which are then windowed using a 1/16th overlap square-root Hann window. The dynamic dictionary (DD), which is essentially an adaptive VQ subsystem, then eliminates signal redundancy. A dictionary of $N\times 1$ codewords is searched for the vector perceptually closest to the input

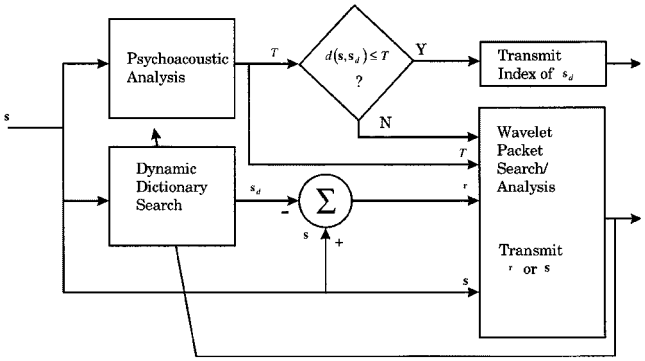


Fig. 29. Dynamic dictionary/optimal wavelet packet encoder (after [157]).

vector. 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 filter bank. A masking threshold, obtained as in [134], is assumed constant within each subband and then used to compute a perceptual bit allocation. The encoder transmits the particular combination of DD and WP information that minimizes the bit rate while maintaining perceptual quality.

This algorithm is unique in that it contains the first reported application of adapted WP analysis to perceptual 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 adaptation is "global" in the sense that the same analysis wavelet is applied to the entire decomposition. The authors reached several useful conclusions regarding the optimal compact support (K-coefficient) wavelet basis when selecting from among the Daubechies orthogonal wavelet bases [159, Proposition 4.5, p. 977]. First, optimization produced average bit-rate savings dependent on filter length of up to 15%. Second, it is not necessary to search exhaustively the space of all wavelets for a particular value of K. The search can be constrained to wavelets with $K/2$ vanishing moments with minimal impact on bit rate. Third, larger K, i.e., more taps, and deeper decomposition trees, tended to

yield better results. As far as quality is concerned, subjective tests showed that the algorithm produced transparent quality for certain test material including drums, pop, violin with orchestra, and clarinet. Subjects detected differences, however, for the castanets and piano sequences. These difficulties arise, respectively, because of inadequate pre-echo control, and inefficient modeling of steady sinusoids. 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 improved quality in the range of 55–63 kb/s, as well as a real-time implementation of on two TMS320C31 devices [160]. Other 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.

2) Scalable DWPT Coder with Adaptive Tree Structure: Srinivasan and Jamieson proposed a WP-based audio coding scheme [161], [162] in which a signal-specific perceptual best basis is constructed by adapting the WP tree structure on each frame such that perceptual entropy and, ultimately, the bit rate are minimized. While the tree structure is signal adaptive, the analysis filters are time invariant and obtained from the family of spline-based biorthogonal wavelets [137]. The algorithm (Fig. 30) is also unique in the sense that it incorporates mechanisms for both bit-rate and complexity scaling. Before the tree adaptation process can commence for a given frame, a set of 63 masking thresholds corresponding to a set of threshold frequency partitions roughly 1/3 Bark wide is obtained from the ISO/IEC MPEG-1 psychoacoustic model recommendation 2 [17]. Of course, depending upon the WP tree, the subbands may or may not align with the threshold partitions. For any particular WP tree, the associated bit rate (cost) is computed by extracting the minimum masking thresholds from each subband and then allocating sufficient bits to guarantee that the quantization noise in each band does not exceed the minimum threshold. The objective of the tree adaptation process, therefore, is to construct a minimum cost subband decomposition by maximizing the minimum masking threshold in every subband. In [161], a complexity-constrained tree adaptation procedure is shown to yield a basis requiring the fewest bits for perceptually transparent coding for a given complexity and temporal resolution. Shapiro's zerotree algorithm [163] is iteratively applied to quantize the coefficients and exploit remaining temporal correlations until the perceptual rate-distortion criteria are satisfied. For informal listening tests over coded program material that included violin, violin/viola, flute, sitar, vocals/orchestra, and sax, the coded outputs at rates in the vicinity of 45 kb/s were reported to be indistinguishable from the originals with the exceptions of the flute and sax. Software is available from the authors' Web site [161]. We note that other researchers have also reported recently on similar strategies for signal-adaptive WP analysis of audio. For example, perceptual metrics for WP tree adaptation were investigated in [164] and [165].

3) DWPT Coder with Globally Adapted General Analysis Wavelet: Srinivasan and Jamieson [161] demonstrated

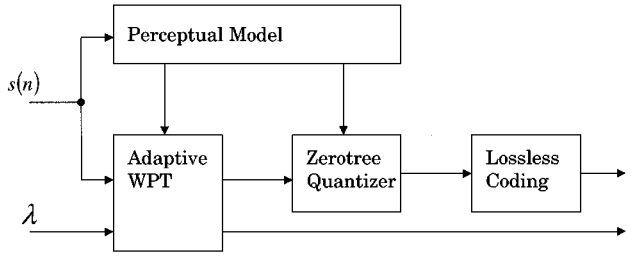


Fig. 30. Masking-threshold adapted WP audio coder [161].

the advantages of a masking threshold adapted WP tree with a time-invariant analysis wavelet. On the other hand, Sinha and Tewfik [158] used a time-invariant WP tree but a globally adapted analysis wavelet to demonstrate that there exists a signal-specific "best" wavelet basis in terms of perceptual coding gain for a particular number of filter taps. The basis optimization in [158], however, was restricted to Daubechies' wavelets. Recent research has attempted to identify which wavelet properties portend an optimal basis, as well as to consider basis optimization over a broader class of wavelets. In an effort to identify the "best" filter, Philippe *et al.* measured the impact on perceptual coding gain of wavelet regularity, AR(1) coding gain, and filter bank frequency selectivity [166], [167]. The study compared performance among orthogonal Rioul [168], orthogonal Onno [169], and the biorthogonal wavelets of [170] in a WP coding scheme that had essentially the same time-invariant critical band WP decomposition tree as [158]. Using filters of lengths varying between 4–120 taps, minimum bit rates required for transparent coding in accordance with the usual perceptual subband bit allocations were measured for each wavelet. For a given filter length, the results suggested that neither regularity nor frequency selectivity mattered significantly. On the other hand, the minimum bit rate required for transparent coding was shown to decrease with increasing analysis filter AR(1) coding gain, leading the authors to conclude that AR(1) coding gain is a legitimate criterion for WP filter selection in perceptual coding schemes.

4) DWPT Coder with Adaptive Tree Structure and Locally Adapted Analysis Wavelet: Phillipe *et al.* [171] measured the perceptual coding gain associated with optimization of the WP analysis filters at every node in the tree, as well as optimization of the tree structure. In one experiment, the WP tree structure was fixed, and then optimal filters were selected for each tree node (local adaptation) such that the bit rate required for transparent coding was minimized. Simulated annealing [172] was used to solve the discrete optimization problem posed by a search space containing 300 filters of varying lengths from the Daubechies [137], Onno [169], Smith–Barnwell [173], Rioul [168], and Akansu–Caglar [174] families. The filters selected by simulated annealing were used in another set of experiments on tree structure optimization. For a fixed tree, the filter adaptation experiments yielded several noteworthy results. First, a nominal bit-rate reduction of 3% was realized for Onno's filters (66.5 kb/s) relative to Daubechies' filters (68 kb/s). Second, simulated annealing over the search space of

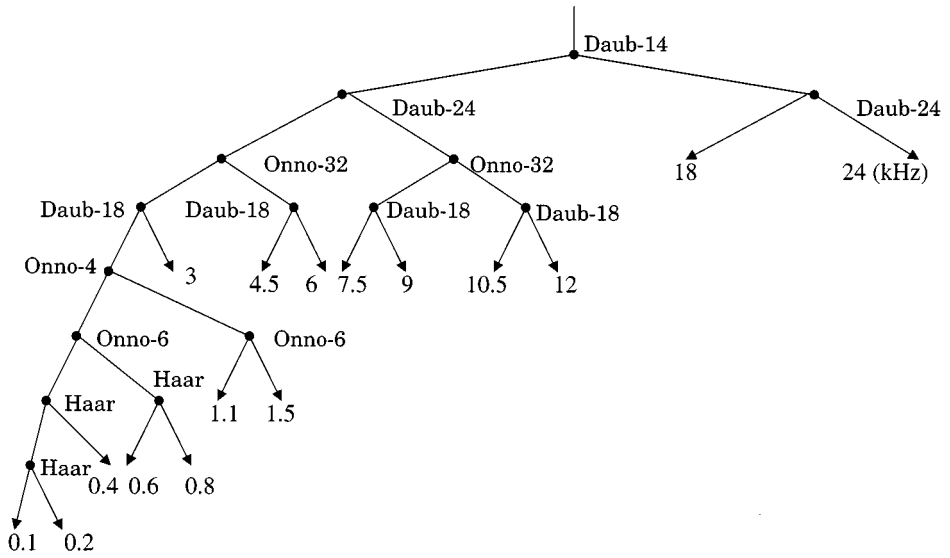


Fig. 31. Wavelet packet analysis filter bank optimized for minimum bit rate, used in MMPE experiments.

300 filters yielded a nominal 1% bit-rate reduction (66 kb/s) relative to the Onno-only case. Finally, longer filter bank delay, i.e., longer analysis filters, yielded lower bit rates. For low-delay applications, however, a seven-fold delay reduction from 700 down to only 100 samples is realized at the cost of only a 10% increase in bit rate. Additional results were reported recently in [175].

5) DWPT Coder with Perceptually Optimized Synthesis Wavelets: Recent research has shown that reconstruction distortion can be minimized in the mean square sense (MMSE) by relaxing PR constraints and tuning the synthesis filters [176]–[182]. Naturally, mean square error minimization is of limited value for subband audio coders. As a result, Gosse *et al.* [183], [184] extended [181] to minimize a mean perceptual error (MMPE) rather than MMSE. A mean perceptual error (MPE) was evaluated at the PR filter bank output in terms of a unique JND measure [185]. Then, an MMPE filter tuning algorithm derived from [181] was applied, and performance was evaluated in terms of a perceptual objective measure [186]. Using the DWPT structure shown in Fig. 31, the authors reported improvement over the PR case, and concluded that further investigation is required to better characterize the costs and benefits of MMPE tuning in a time-varying scenario.

E. Hybrid Harmonic/Wavelet Decompositions

Although the WP coder improvements reported in [160] addressed pre-echo control problems evident in [158], they did not rectify the coder's inadequate performance for harmonic signals such as the piano test sequence. This is in part because the low-order FIR analysis filters typically employed in a WP decomposition are characterized by poor frequency selectivity, and therefore wavelet bases tend not to provide compact representations for strongly sinusoidal signals. On the other hand, wavelet decompositions provide some control over time resolution properties, leading to efficient representations of transient signals. These considerations have inspired several researchers to investigate hybrid coders.

1) Hybrid Sinusoidal/Classical DWPT Coder: Hamdy *et al.* developed a novel hybrid coder [187] designed to exploit the efficiencies of both harmonic and wavelet signal representations. For each analysis frame, the encoder (Fig. 32) 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, Thomson's analysis model [188] is applied to extract sinusoidal parameters for each input frame. Harmonic synthesis using the McAulay and Quatieri reconstruction algorithm [189] 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 filter bank. Edge-detection processing identifies and removes transients in low-frequency subbands. Without transients, the residual WP coefficients 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 are concerned, sinusoidal frequencies are quantized with sufficient precision to satisfy just-noticeable-differences in frequency (JNDF). Sinusoidal amplitudes are quantized and encoded in accordance with a masked threshold estimate. Sinusoidal phases are uniformly quantized on the interval $[-\pi, \pi]$. As for quantization and encoding of WP parameters, all coefficients below 11 kHz are encoded as in [371]. Above 11 kHz, however, parametric representations are utilized. Transients are represented in terms of a binary edge mask, while noise components are represented in terms of means, variances, and decay constants. The

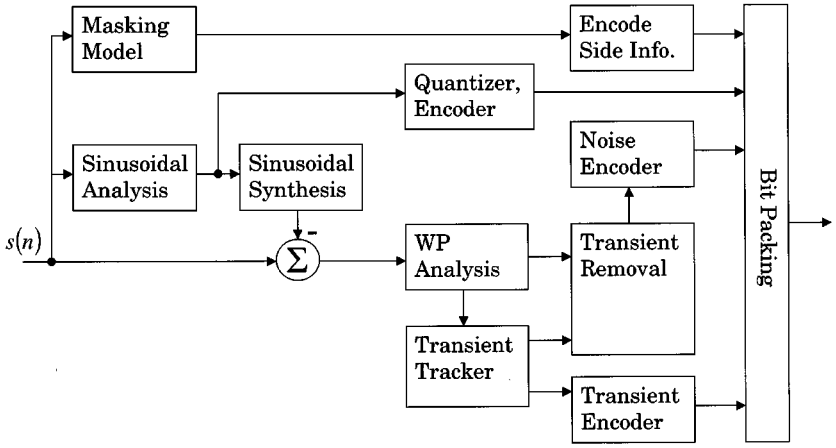


Fig. 32. Hybrid sinusoidal/wavelet encoder (after [187]).

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 kb/s [190].

2) Hybrid Sinusoidal/M-Band DWPT Coder: Boland and Deriche [138] reported on an experimental sinusoidal-wavelet hybrid audio codec with high-level architecture very similar to [187] but with low-level differences in the sinusoidal and wavelet analysis blocks. In particular, for harmonic analysis the proposed algorithm replaces Thomson's method used in [187] with a combination of total least squares linear prediction (TLS-LP) and Prony's method. Then, in the harmonic residual wavelet decomposition block, the proposed method replaces the usual DWT cascade of two-band QMF sections with a cascade of four-band QMF sections. In the wavelet analysis section, the harmonic residual $r(n)$ is decomposed such that critical bandwidths are roughly approximated using a three-level cascade of four-band analysis filters (i.e., ten subbands) designed according to the M-band technique in [191]. After subjective listening comparisons between the proposed scheme at 60–70 kb/s and MPEG-1, Layer III at 64 kb/s on 12 SQAM CD [192] source items, the authors reported indistinguishable quality for "acoustic guitar," "Eddie Rabbit," "castanets," and "female speech."

3) Hybrid Sinusoidal/DWPT Coder with Tree Structure Adaptation (ARCO): Pena *et al.* [139] have reported on the "Adaptive Resolution Codec" (ARCO). This algorithm employs a two-stage hybrid tonal-WP analysis section architecturally similar to both [187] and [138]. ARCO introduced several novelties in the segmentation, psychoacoustic analysis, and WP analysis blocks. In an effort to match the time-frequency analysis resolution to the signal properties, ARCO includes a subframing scheme that makes use of both time and frequency block clustering to determine optimal analysis frame lengths [193]. The ARCO psychoacoustic model resembles ISO/IEC MPEG-1 model recommendation 1 [17], with some enhancements. Tonality labeling is based on [194], and noise maskers are segregated into narrow-band and wide-band subclasses. Wide-band noise maskers have frequency-dependent excitation patterns. The ARCO WP decomposition procedure optimizes both the tree structure, as in [161], and filter selections, as in [158] and [171]. ARCO essentially arranges the subbands such that the corresponding set of idealized brickwall rectangular filters having amplitude equal to the height of the minimum masking threshold in the each band matches as closely as possible the shape of the masking threshold. Bits are allocated in each subband to satisfy the minimum masking threshold A_k. The ARCO bit allocation strategy [195] achieves fast convergence to a desired bit rate by shifting the masking threshold up or down. Another unique property of ARCO is its set of high-level "cognitive rules" that seek to minimize the objectionable distortion when insufficient bits are available to guarantee transparent coding [196]. Finally, it is interesting to note that researchers developing ARCO recently replaced the hybrid sinusoidal-WP analysis filter bank with a novel multiresolution MDCT-based filter bank. In [197], Casal *et al.* developed a "Multi-Transform" (MT) that retains the lapped properties of the MDCT but creates a nonuniform time-frequency tiling by transforming back into time the high-frequency MDCT components in L-sample blocks. The proposed MT is characterized by high resolution in frequency in the low subbands and high resolution in time at the high frequencies.

F. Signal-Adaptive, Nonuniform Filter Bank (NUFB) Decompositions

The most popular method for realizing nonuniform frequency subbands is to cascade uniform filters in an unbalanced tree structure, as with, for example, the DWPT. For a given impulse response length, however, cascade structures in general produce poor channel isolation. Recent advances in modulated filter bank design methodologies (e.g., [198]) have made tractable direct form near perfect reconstruction nonuniform designs, which are critically sampled. We next consider subband coders that employ signal-adaptive nonuniform modulated filter banks to approximate the time-frequency analysis properties of the auditory system more effectively than the other subband coders. Beyond the algorithms addressed below, we note that other investigators have proposed nonuniform filter bank coding techniques,

which address redundancy reduction utilizing lattice [199] and bidimensional VQ schemes [200].

1) Switched Nonuniform Filter Bank Cascade: Princen and Johnston developed a CD-quality coder based upon a signal-adaptive filter bank [201] for which they reported quality better than the sophisticated MPEG-1 Layer III algorithm at both 48 and 64 kb/s. The analysis filter bank for this coder consists of a two-stage cascade. The first stage is a 48-band nonuniform modulated filter bank split into four uniform-bandwidth sections. There are eight uniform subbands from 0 to 750 Hz, four uniform subbands from 750 to 1500 Hz, 12 uniform subbands from 1.5 to 6 kHz, and 24 uniform subbands from 6 to 24 kHz. The second stage in the cascade optionally decomposes nonuniform bank outputs with on/off switchable banks of finer resolution uniform subbands. During filter bank 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. Uniform PCM is applied to subband samples under the constraint of perceptually masked quantization noise.

2) FV-MLT: Purat and Noll [370] 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 [202]. Second, 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 distortions. In informal subjective tests, the algorithm achieved excellent quality at a bit rate of 55 kb/s.

G. IIR Filter Banks

Although the majority of subband and wavelet audio coding algorithms found in the literature employ banks of perfect reconstruction FIR filters, this does not preclude the possibility of using infinite impulse response (IIR) filter banks for the same purpose. Compared to FIR filters, IIR filters are able to achieve similar magnitude response characteristics with reduced filter orders, and hence with reduced complexity. In the multiband case, IIR filter banks also offer complexity advantages over FIR filter banks. Enhanced performance, however, comes at the expense of an increased construction and implementation effort for IIR filter banks. Creusere and Mitra constructed a template subband audio coding system modeled after [366] to compare performance and to study the tradeoffs involved when choosing between FIR and IIR filter banks for the audio coding application [203]. In the study, two IIR and two FIR coding schemes were constructed from the template using a structured all-pass filter bank, a parallel allpass filter bank, a tree-structured QMF bank, and a polyphase quadrature filter bank.

VI. Sinusoidal Coders

Although sinusoidal signal models have been applied successfully in speech coding [204], [205], [189], [212] and music synthesis applications [214], there was until recently relatively little work reported on perceptual audio coding using sinusoidal signal models. The existing sinusoidal coders were developed in a speech coding context, and tended to minimize MSE. Perceptual properties were introduced later [139], [206], [207], [211]. This section is concerned with perceptual coding algorithms based on purely sinusoidal or hybrid sinusoidal signal models. The advent of MPEG-4 standardization established new research goals for high-quality coding of general audio signals at bit rates in the range of 6–24 kb/s, rates that had previously been reserved for speech-specific coding algorithms. The problem addressed in the MPEG-4 research was to achieve low rates while eliminating the source-system paradigm that characterizes most speech coders. In experiments reported as part of the MPEG-4 standardization effort, it was determined that sinusoidal coding is capable of achieving good quality at low rates without being constrained by a restrictive source model. Furthermore, unlike CELP and other classical low-rate speech coding models, the parametric sinusoidal coding is amenable in a straightforward manner to pitch and time-scale modification at the decoder. This section describes sinusoidal algorithms recently proposed for low-rate audio coding using perceptual properties, including the Analysis/Synthesis Audio Codec (ASAC), enhanced ASAC, and FM ASAC. Some of these methodologies have been adopted as a part of the MPEG-4 standardization (Section VIII). Additionally, outside of the MPEG-4 standardization framework, the recent emergence of Internet-based streaming audio has motivated considerable research on the application of sinusoidal signal models to high-quality audio coding at low bit rates. For example, Levine and Smith developed a hybrid sinusoidal-filter bank coding scheme that achieves very high quality at rates in the vicinity of 32 kb/s [206], [208], [209].

A. Analysis/Synthesis Audio Codec

The sinusoidal ASAC for robust coding of general audio signals at rates between 6 and 24 kb/s was developed by Edler *et al.* at the University of Hannover and proposed for MPEG-4 standardization [210] in 1995. An enhanced ASAC proposal later appeared in [211]. Initially, ASAC segments input audio into analysis frames over which the signal is assumed to be nearly stationary. Sinusoidal synthesis parameters are then extracted according to perceptual criteria, quantized, encoded, and transmitted to the decoder for synthesis. The algorithm distributes synthesis parameters across basic and enhanced bitstreams to allow scalable output quality at bit rates of 6 and 24 kb/s. Architecturally, the ASAC scheme (Fig. 33) consists of a preanalysis block for window selection and envelope extraction, a sinusoidal analysis-by-synthesis parameter estimation block, a perceptual model, and a quantization and coding block. Although it bears similarities to sinusoidal speech coding [189], [212], [213] and music synthesis [214] algorithms that have been available for some time, the ASAC coder also incorporates some new techniques. In particular, whereas previous sinusoidal coders emphasized waveform matching

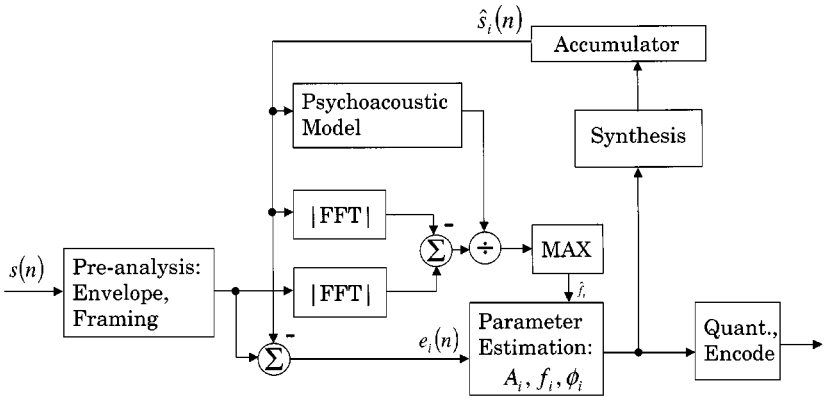


Fig. 33. ASAC encoder (after [216]).

by minimizing reconstruction error norms such as the mean square error, ASAC disregards classical error minimization criteria and instead selects sinusoids in decreasing order of *perceptual* importance by means of an iterative analysis-by-synthesis loop. The perceptual significance of each component sinusoid is judged with respect to the masking power of the synthesis signal, which is determined by a simplified version of the psychoacoustic model [215]. The iterative analysis-by-synthesis block [216] estimates one at a time the parameters of the ith individual constituent sinusoid or partial, and every iteration identifies the most perceptually significant sinusoid remaining in the synthesis residual, $e_i(n) = s(n) - \hat{s}_i(n)$, and adds it to the synthetic output, $\hat{s}_i(n)$. Perceptual significance is assessed by comparing the synthesis residual against the masked threshold associated with the current synthetic output and choosing the residual sinusoid with the largest suprathreshold margin. The loop repeats until the bit budget is exhausted. When compared to standard speech codecs at similar bit rates, the first version of ASAC [210] reportedly offered improved quality for nonharmonic tonal signals such as spectrally complex music, similar quality for single instruments, and impaired quality for clean speech [217]. The later ASAC [211] was improved for certain signals [218].

B. Harmonic and Individual Lines Plus Noise Coder

The ASAC algorithm outperformed speech-specific algorithms at the same bit rate in subjective tests for some test signals, particularly spectrally complex music characterized by large numbers of nonharmonically related sinusoids. The original ASAC, however, failed to match speech codec performance for other test signals such as clean speech. As a result, the ASAC core was embedded in an enhanced algorithm [219] intended to better match the coder's signal model with diverse input signal characteristics. In research proposed as part of an MPEG-4 "core experiment" [220], Purnhagen *et al.* at the University of Hannover developed in conjunction with Deutsche Telekom Berkom an "object-based" algorithm. In this approach, harmonic sinusoid, individual sinusoid, and colored noise objects could be combined in a hybrid source model to create a parametric signal representation. The enhanced algorithm, known as the "Harmonic and Individual Lines Plus Noise" (HILN), is architecturally very similar to the original ASAC, with some modifications (Fig. 34). The iterative analysis–synthesis block is extended to include a cascade of analysis stages for each of the available object types. In the enhanced analysis–synthesis system, harmonic analysis is applied first, followed by individual spectral line analysis, followed by shaped noise modeling of the two-stage residual. Results from subjective listening tests at 6 kb/s showed significant improvements for HILN over ASAC, particularly for the most critical test items that had previously generated the most objectionable ASAC artifacts [221]. Compared to HILN, CELP speech codecs are still able to represent more efficiently clean speech at low rates, and "time-frequency" codecs are able to encode more efficiently general audio at rates above 32 kb/s. Nevertheless, the HILN improvements relative to ASAC inspired the MPEG-4 committee to incorporate HILN into the MPEG-4 committee draft as the recommended low-rate parametric audio coder [222]. The HILN algorithm was recently deployed in a scalable low-rate Internet streaming audio scheme [223].

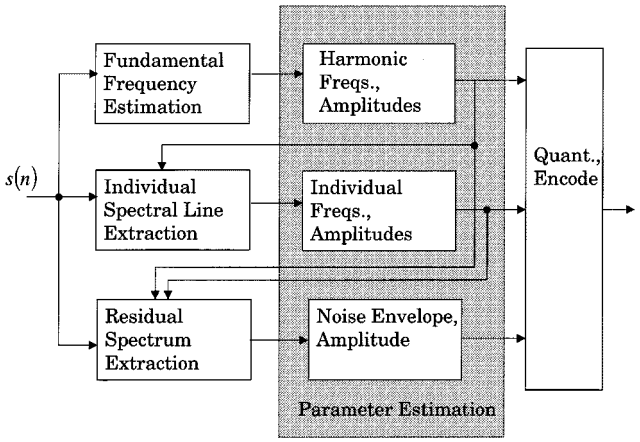


Fig. 34. HILN encoder (after [219]).

C. FM Synthesis

The HILN algorithm seeks to optimize coding efficiency by making combined use of three distinct source models. Although the HILN harmonic sinusoid object has been

shown to facilitate increased coding gain for certain signals, it is possible that other object types may offer opportunities for greater efficiency when representing spectrally complex harmonic signals. This notion motivated a recent investigation into the use of frequency modulation (FM) synthesis techniques [224] in low-rate sinusoidal audio coding for harmonically structured single instrument sounds [225]. FM synthesis offers advantages over other harmonic coding methods (e.g., [216], [226]) because of its ability to model with relatively few parameters harmonic signals that have many partials. In the simplest FM synthesis, for example, the frequency of a sine wave (carrier) is modulated by another sine wave (modulator) to generate a complex waveform with spectral characteristics that depend on a modulation index and the parameters of the two sine waves. In continuous time, the FM signal is given by

$$s(t) = A \sin[2\pi f_c t + I \sin(2\pi f_m t)] \tag{49}$$

where

A	amplitude;
f_c	carrier frequency;
f_m	modulation frequency;
I	modulation index;
t	time index.

The associated Fourier series representation is

$$s(t) = \sum_{k=-\infty}^{\infty} J_k(I) \sin(2\pi f_c t + 2\pi k f_m t) \tag{50}$$

where $J_k(I)$ is the Bessel function of the first kind. It can be seen from (50) that a large number of harmonic partials can be generated (Fig. 35) by controlling only three parameters per FM "operator." One can observe that the fundamental and harmonic frequencies are determined by f_c and f_m, and that the harmonic partial amplitudes are controlled by the modulation index I. The Bessel envelope, moreover, essentially determines the FM spectral bandwidth. Example harmonic FM spectra for a unit amplitude 200-Hz carrier are given in Fig. 35 for modulation indexes of one [Fig. 35(a)] and 15 [Fig. 35(b)]. While both examples have identical harmonic structure, the amplitude envelopes and bandwidths differ markedly as a function of the index I. Clearly, the central issue in making effective use of the FM technique for signal modeling is parameter estimation accuracy.

Winduratna proposed an FM synthesis audio coding scheme in which the outputs of parallel FM "operators" are combined to model a single instrument sound. The algorithm (Fig. 36) works as follows. First, the preanalysis block segments input audio into analysis frames and then extracts parameters for a set of individual spectral lines, as in [216]. Next, the preanalysis identifies a harmonic structure by maximizing an objective function [225]. Given a fundamental frequency estimate from the preanalysis f_0, the iterative parameter extraction loop estimates the parameters of individual FM operators and accumulates their contributions until the composite spectrum closely resembles the original. Perceptual closeness is judged to be adequate when the absolute original minus synthetic harmonic difference

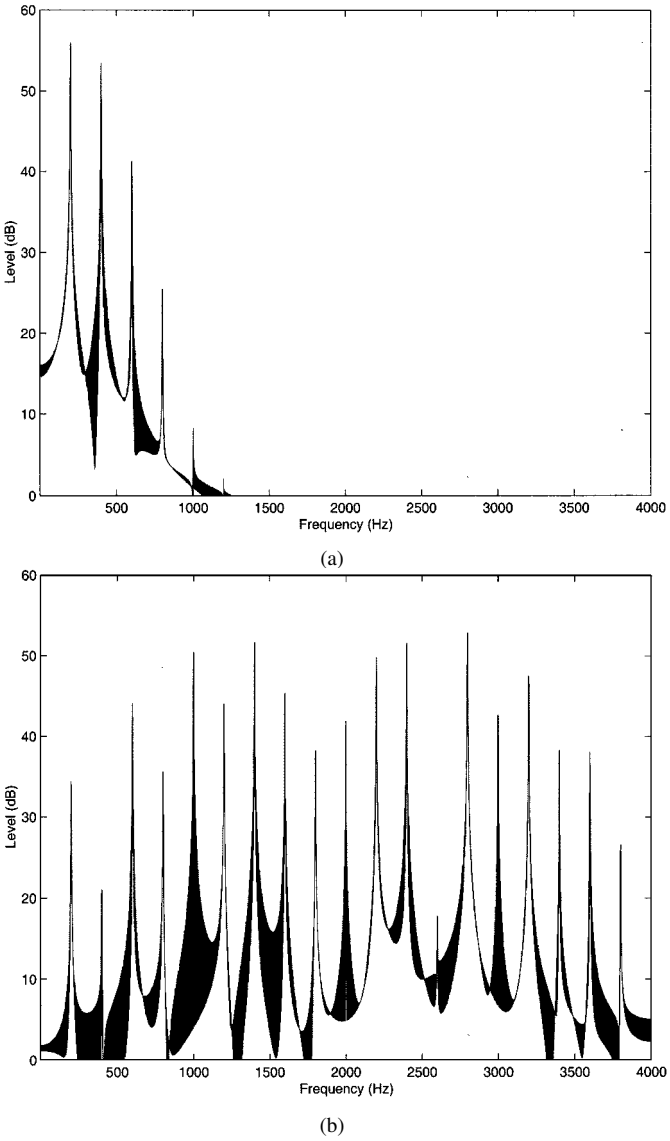


Fig. 35. Harmonic FM spectra, $f_c = f_m = 200$ Hz, with (a) $I = 1$ and (b) $I = 15$.

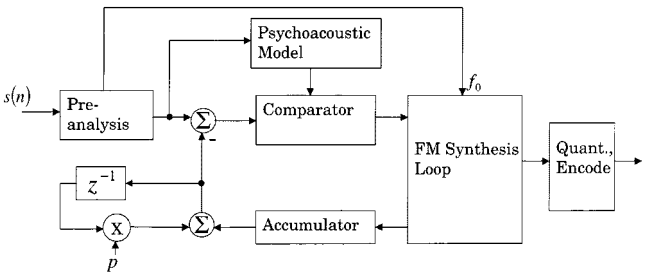


Fig. 36. FM synthesis coding scheme (after [225]).

spectrum is below the masked threshold [215]. During each loop iteration, error minimizing values for the current operator are determined by means of an exhaustive search. The loop repeats and additional operators are synthesized until the error spectrum is below the masked threshold. The FM coding scheme was shown to efficiently represent single instrument sounds at bit rates between 2.1–4.8 kb/s. Using a 24-ms analysis window, for example, one critical male

speech item was encoded at 21.2 kb/s using FM synthesis compared to 45 kb/s for ASAC [225], with similar output quality. Despite estimation difficulties for signals with more than one fundamental, e.g., polyphonic music, the high efficiency of the FM synthesis technique makes it a likely candidate for future inclusion in object-based algorithms such as HILN.

D. Hybrid Sinusoidal Coders

Whereas the waveform-preserving perceptual transform (Section IV) and subband (Section IV) coders tend to target transparent quality at bit rates between 32–128 kb/s per channel, the sinusoidal coders proposed thus far in the literature have concentrated on very low-rate applications between 2–16 kb/s. Rather than transparent quality, these algorithms have emphasized source robustness, i.e., the ability to deal with general audio at low rates without constraining source model dependence. The current low-rate sinusoidal algorithms (ASAC, HILN, etc.) represent the perceptually significant portions of the magnitude spectrum from the original signal without explicitly treating the phase spectrum. As a result, perceptually transparent coding is typically not achieved with these algorithms. It is generally agreed that different state-of-the-art coding techniques perform most efficiently in terms of output quality achieved for a given bit rate. In particular, CELP speech algorithms offer the best performance for clean speech below 16 kpbs, parametric sinusoidal techniques perform best for general audio between 16–32 kb/s, and so-called time-frequency audio codecs tend to offer the best performance at rates above 32 kb/s. Designers of comprehensive bit-rate scalable coding systems, therefore, must decide whether to cascade multiple stages of fundamentally different coder architectures with each stage operating on residual signal from the previous stage, or alternatively to "simulcast" independent bitstreams from different coder architectures and then select an appropriate decoder at the receiver. In fact, some experimental work performed in the context of MPEG-4 standardization demonstrated that a cascaded, hybrid sinusoidal/time-frequency coder can not only meet but in some cases even exceed the output quality achieved by the time-frequency (transform) coder alone at the same bit rate for certain critical test signals [227]. Issues critical to cascading successfully a parametric sinusoidal coder with a transform-based time-frequency coder are addressed in [228]. It was earlier noted that CELP speech algorithms typically outperform the parametric sinusoidal coders for clean speech inputs at rates below 16 kb/s. There is some uncertainty, however, as to which class of algorithm is best suited when both speech and music are present. A hybrid scheme (Fig. 37) intended to outperform CELP/parametric "simulcast" for speech/music mixtures was proposed in [228]. As expected, the hybrid structure was reported to outperform simulcast configurations only when the voice signal was dominant [228]. Quality degradations were reported for mixtures containing dominant musical signals. In the future, hybrid structures of this type will benefit from emerging techniques in speech/music discrimination (e.g., [229], [230]). As observed by Edler, on the other hand, future audio coding research is also quite likely to focus on automatic decomposition of complex input signals into components for which individual coding is more efficient than direct coding of the mixture [231] using hybrid structures. Advances in sound separation and auditory scene analysis [232], [233] techniques will eventually make the automated decomposition process viable.

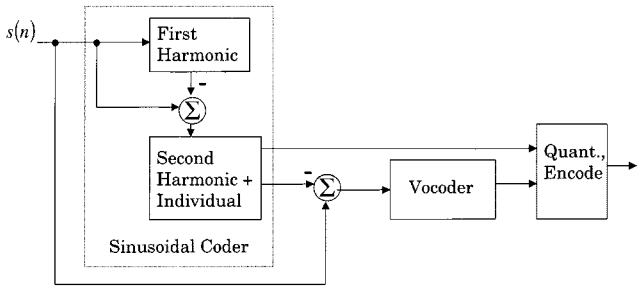


Fig. 37. Hybrid sinusoidal/vocoder (after [228]).

VII. Linear-Prediction-Based Coders

Although other methodologies have been the focus of attention in perceptual audio coding research, a few CD-quality coders based on a source-system model and linear prediction have also been reported to achieve transparent or near transparent quality with bit rates ranging between 64–128 kb/s. With the exception of TwinVQ [128], however, the LP audio codecs have primarily remained within the experimental domain. In light of the recent trend toward hybrid speech and audio coding at rates below 16 kb/s, it is useful to consider existing LP techniques in audio coding. It was observed in formal listening tests during MPEG-4 standardization, for example, that at certain low rates, the best choice of signal model depends upon the source material. In particular, a CELP coder outperforms a sinusoidal coder for speech, but the sinusoidal coder outperforms the CELP coder for music. It is conceivable that a more efficient future hybrid algorithm will capitalize on the strengths of both signal models in a single coder. The benefits of perceptual LP codecs in this scenario as yet have been largely unexplored. In spite of the fact that the LP analysis–synthesis framework is central to modern speech coding algorithms [234], it has received relatively little attention in the audio coding literature or standards. One reason is that the LP coders are not well suited to the task of modeling the nearly sinusoidal components present in steady-state audio signals. These elements create sharp peaks in the spectral envelope, which often in the presence of quantization noise lead to LP synthesis filter instabilities. Another reason for the lack of interest is that the source-system represented by the LP analysis–synthesis framework does not necessarily model any of the physical mechanisms that generate audio signals. The correspondence between the LP analysis–synthesis and the source-system speech production model has been a primary reason for its success in speech applications. Whether or not LP analysis–synthesis is well suited to modeling

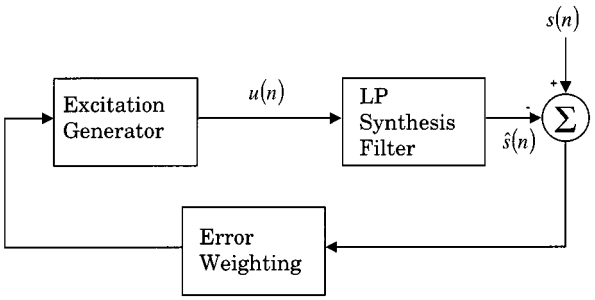


Fig. 38. Multipulse excitation model used in [235].

audio is highly signal-dependent. Nevertheless, several LP algorithms have been successfully applied to CD-quality audio. This section considers some examples of LP-based audio codecs. In addition, the section examines a novel coder based on frequency-warped LP that has potential for reduced complexity by eliminating the explicit perceptual model.

A. Multipulse Excitation

Singhal at Bell Labs [235] reported that analysis-by-synthesis multipulse excitation of sufficient pulse density can be applied to correct for LP envelope errors introduced by bandwidth expansion and quantization (Fig. 38). This algorithm uses a twenty-fourth-order LPC synthesis filter while optimizing pulse positions and amplitudes to minimize perceptually weighted reconstruction errors. Singhal determined that densities of approximately one pulse per four output samples of each excitation subframe are required to achieve near transparent quality. Spectral coefficients are transformed to inverse sine reflection coefficients, then differentially encoded and quantized using pdf-optimized Max quantizers. Entropy (Huffman) codes are also used. Pulse locations are differentially encoded relative to the location of the first pulse. Pulse amplitudes are fractionally encoded relative to the largest pulse and then quantized using a Max quantizer. The proposed MPLPC audio coder achieved output SNR's of 35–40 dB at a bit rate of 128 kb/s. Other MPLPC audio coders have also been proposed [236], including a scheme based on MPLPC in conjunction with the discrete wavelet transform [147].

B. Discrete Wavelet Excitation Coding

While the most successful speech codecs nowadays use some form of closed-loop time-domain analysis-by-synthesis such as MPLPC, high-performance LP-based perceptual audio coding has been realized with alternative frequency-domain excitation models. For instance, Boland and Deriche reported output quality comparable to MPEG-1, Layer II at 128 kb/s for an LPC audio coder operating at 96 kb/s [237] in which the prediction residual was transform coded using a three-level DWT based on a four-band uniform filter bank. At each level of the DWT, the lowest subband of the previous level was decomposed into four uniform bands. This ten-band nonuniform structure was intended to mimic critical bandwidths to a certain extent. A perceptual bit allocation according to MPEG-1, psychoacoustic model 2 was applied to the transform coefficients.

C. Sinusoidal Excitation Coding

Still other frequency-domain excitation models are possible. Excitation sequences modeled as a sum of sinusoids were investigated [238] in order to capitalize on the experimentally observed tendency of the prediction residuals for high-fidelity audio to be spectrally impulsive rather than flat. In coding experiments using 32-kHz-sampled input audio, subjective and objective quality improvements relative to the MPLPC coders were reported for the sinusoidal excitation schemes, with high-quality output audio reported at 72 kb/s. In the experiments [239], a set of tenth-order LP coefficients is estimated on 9.4-ms analysis frames and split-vector quantized using 24 bits. Then, the prediction residual is analyzed and sinusoidal parameters are estimated for the seven best out of a candidate set of 13 sinusoids for each of six subframes. The masked threshold is estimated and used to form a time-varying bit allocation for the amplitudes, frequencies, and phases on each subframe. Given a frame allocation of 675, a total of 573, 78, and 24 bits, respectively, are allocated to the sinusoidal, bit allocation side information, and LP coefficients. In conjunction with the usage of a masking-threshold adapted weighting filter, the sinusoidal excitation scheme was also reported to deliver improved quality relative to MPEG-1, Layer I at a bit rate of 96 kb/s [238] for selected test material, including piano, horn, and drum.

D. Frequency Warped LP

Beyond the performance improvements realized through the use of different excitation models, there has been some interest in warping the frequency axis prior to performing LP analysis to effectively provide better resolution at some frequencies than at others. In the context of perceptual coding, it is naturally of interest to achieve a Bark-scale warping. Frequency axis warping to achieve nonuniform FFT resolution was first introduced by Oppenheim *et al.* [240], [241] using a network of cascaded first-order all-pass sections for frequency warping of the signal, followed by a standard FFT. The idea was later extended to warped linear prediction (WLP) by Strube [242], and was ultimately applied in an ADPCM codec [243]. Cascaded First-order all-pass sections were used to warp the signal, and then the LP autocorrelation analysis was performed on the warped autocorrelation sequence. In this scenario, a single-parameter warping of the frequency axis can be introduced into the LP analysis by replacing the delay elements in the FIR analysis filter with all-pass sections, i.e., by replacing the complex variable z with a filter $H(z)$ of the form

$$H(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}}. \tag{51}$$

Thus, the predicted sample value is not produced from a combination of past samples, but rather from the samples of a warped signal. In fact, it has been shown [244], [405] that selecting the value of 0.723 for the parameter λ leads to a

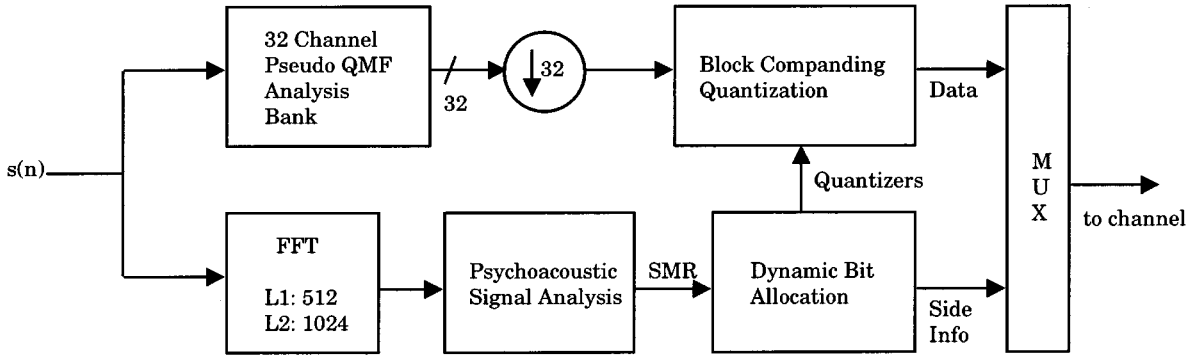


Fig. 39. ISO/IEC 11172-3 (MPEG-1) layer I/II encoder.

frequency warp that approximates well the Bark frequency scale. A WLP-based audio codec [245] was recently proposed. The inherent Bark frequency resolution of the WLP prediction residual yields a perceptually shaped quantization noise without the use of an explicit perceptual model or time-varying bit allocation. In this system, a fortieth-order WLP synthesis filter is combined with differential encoding of the prediction residual. A fixed rate of 2 bits per sample (88.2 kb/s) is allocated to the residual sequence, and 5 bits per coefficient are allocated to the prediction coefficients on an analysis frame of 800 samples, or 18 ms. This translates to a bit rate of 99.2 kb/s per channel. In objective terms, an auditory error measure showed considerable improvement for the WLP coding error in comparison to a conventional LP coding error when the same number of bits was allocated to the prediction residuals. Subjectively, the algorithm was reported to achieve transparent quality for some material, but it also had difficulty with transients at the frame boundaries. The algorithm was later extended to handle stereophonic signals [246] by forming a complex-valued representation of the two channels and then using WLP for complex signals (CWLP). Less than CD quality was reported at a rate of 128 kb/s for 44.1-kHz-sampled source material. It was suggested that significant quality improvement could be realized for the WLPC audio coder by improving the excitation model to use a closed-loop analysis-by-synthesis procedure such as CELP or a multipulse model [247]. One of the shortcomings of the original WLP coder was inadequate attention to temporal effects. As a result, further experiments were reported [248] in which WLP was combined with TNS to realize additional quality improvement for the complex-signal stereophonic WLP audio coder. Future developments in LP-based audio codecs will continue to appear, particularly in the context of low-rate hybrid coders for both speech and audio.

VIII. Audio Coding Standards

This section gives both high-level descriptions and important details of several international and commercial product audio coding standards, including the ISO/IEC MPEG-1/-2/-4 series, the Dolby AC-2/AC-3, the Sony ATRAC/MiniDisc/SDDS, the Lucent Technologies PAC/EPAC/MPAC, and the Phillips DCC algorithms.

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

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 [249] comprises a flexible hybrid coding technique, which incorporates several methods including subband decomposition, filter bank 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) [250]–[252] 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 was backward compatible with MPEG-1 in the sense that 3/2 channel information transmitted by an MPEG-2 encoder can be correctly decoded for two-channel presentation by an MPEG-1 receiver. The second MPEG-2 standard sacrificed backward MPEG-1 compatibility to eliminate quantization noise unmasking artifacts [253] which are potentially introduced by the forced backward compatibility. Several tutorials on the MPEG-1 [254]–[257] and MPEG-1/2 [30], [31], [75] standards have appeared. MPEG standardization work is continuing, and will eventually lead to very low rates for high fidelity, perhaps reaching as low as 16 kb/s per channel.

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. 39) work as follows. The input signal is first decomposed into 32 critically subsampled subbands using a polyphase realization of a PQMF bank [78] (Section III). The channels are equally spaced such that a 48-kHz input signal is split into 750-Hz subbands, with the subbands decimated 32 : 1. A 511th-order prototype filter was chosen such that the inherent overall PQMF distortion remains below the threshold of audibility. Moreover, the prototype filter was designed for very high sidelobe attenuation (96 dB) to insure that intraband aliasing due to quantization

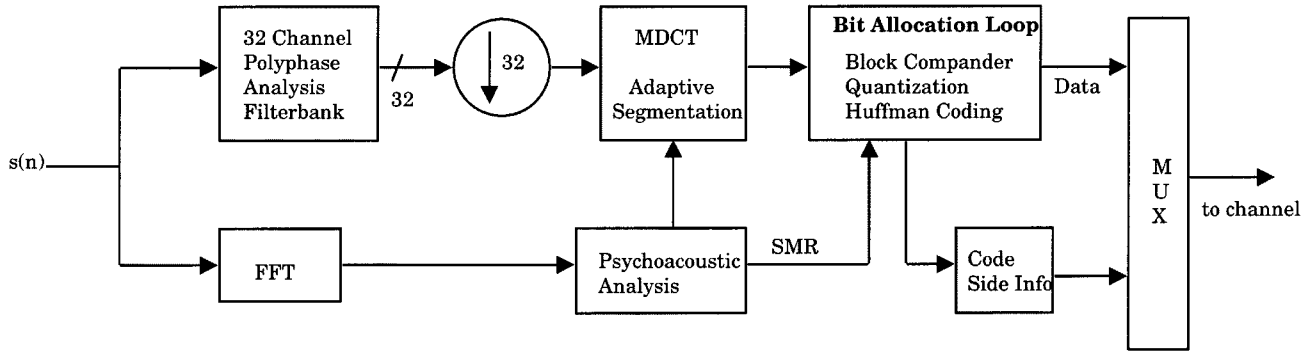


Fig. 40. ISO/IEC 11172-3 (MPEG-1) layer III encoder.

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, scale factors are quantized using 6 bits and quantizer selections are encoded using 4 bits.

1) Layer I: For layer I encoding, decimated subband sequences are quantized and transmitted to the receiver in conjunction with side information, including quantized scale factors and quantizer selections.

2) Layer II: Layer II improves three portions of Layer I in order to realize enhanced output quality and reduce bit rates at the expense of greater complexity and increased delay. In particular, the perceptual model relies upon a higher resolution FFT, the maximum subband quantizer resolution is increased, and scale-factor side information is reduced while exploiting temporal masking by considering properties of three adjacent 12-sample blocks and optionally transmitting one, two, or three scale factors. Average MOS's of 4.7 and 4.8 were reported [30] for monaural layer I and layer II codecs operating at 192 and 128 kb/s, respectively. Averages were computed over a range of test material.

3) Layer III: The layer III MPEG (Fig. 40) architecture achieves performance improvements by adding several important mechanisms on top of the layer I/II foundation. A hybrid filter bank is introduced to increase frequency resolution and thereby better approximate critical band behavior. The hybrid filter bank includes adaptive segmentation to improve pre-echo control. Sophisticated bit allocation and quantization strategies that rely upon nonuniform quantization, analysis-by-synthesis, and entropy coding are introduced to allow reduced bit rates and improved quality. The hybrid filter bank 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. The adaptive MDCT switches between 6–18 points to allow improved pre-echo control. Shorter blocks (4 ms) provide for temporal premasking of pre-echoes during transients; longer blocks during steady-state periods improve coding gain, while also reducing side information and hence bit rates. Bit allocation and quantization of the spectral lines are realized in a nested loop procedure that uses both nonuniform quantization and Huffman coding. The inner loop adjusts the nonuniform 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 [30] for monaural layer II and layer III codecs operating at 64 kb/s.

4) Applications: MPEG-1 has been successful in numerous applications. For example, MPEG-1 Layer III has become the *de facto* standard for transmission and storage of compressed audio for both WWW and handheld media applications (e.g., Diamond RIO). In these applications, the ".MP3" label denotes MPEG-1, Layer III. Note that MPEG-1 audio coding has steadily gained acceptance and ultimately has been deployed in several other large scale systems, including the European digital radio (DBA) or Eureka [359], direct broadcast satellite [360], and digital compact cassette [366]. Recently, moreover, the collaborative European Advanced Communications Technologies and Services (ACTS) program adopted MPEG audio and video as the core compression technologies for the Advanced Television at Low Bitrates And Networked Transmission over Integrated Communication systems (ATLANTIC) project, a system intended to provide functionality for television program production and distribution [258], [259]. The ATLANTIC system has posed new challenges for MPEG deployment such as seamless bitstream (source) switching [260] and robust transcoding (tandem coding). Unfortunately, transcoding is neither guaranteed nor likely to preserve perceptual noise masking [261]. A buried data "MOLE" signal was proposed to mitigate and in some cases eliminate transcoding distortion for cascaded MPEG-1 layer II codecs [262], ideally allowing downstream tandem stages to preserve the original bitstream. The idea behind the MOLE is to apply the same set of quantizers to the same set of data in the downstream codecs as in the original codec. The output bitstream will then be identical to the original

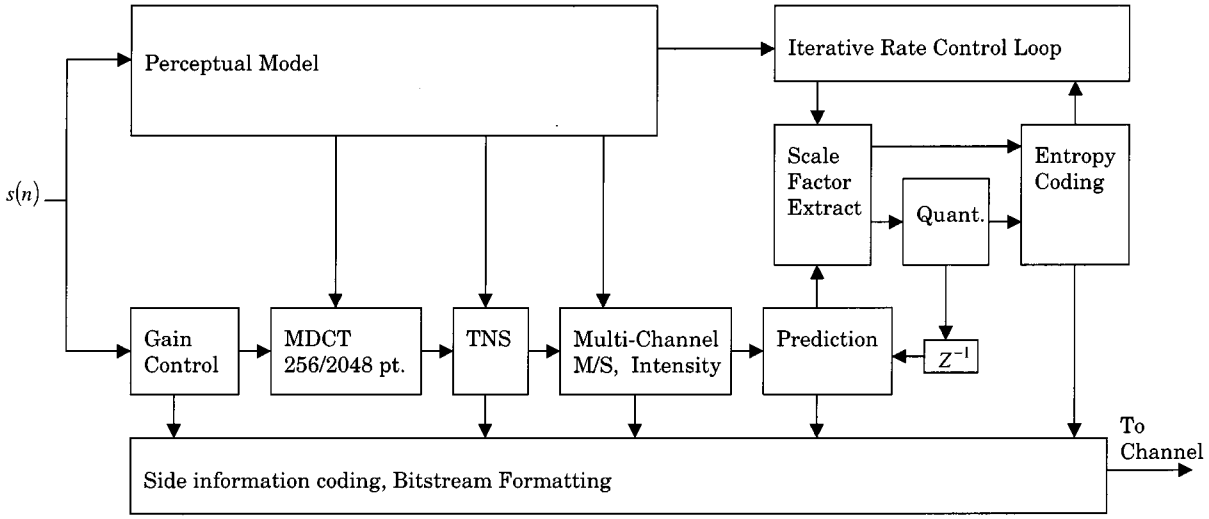


Fig. 41. ISO/IEC IS13818-7 (MPEG-2 NBC/AAC) encoder (after [266]).

bitstream, provided that numerical precision in the analysis filter banks does not bias the data [263].

We will next consider the more recent and in some cases still-evolving MPEG standards for audio, namely, the MPEG-2 AAC and the MPEG-4 algorithms. The discussion will focus primarily upon architectural novelties and differences from MPEG-1.

B. ISO/IEC IS13818-7 (MPEG-2 NBC/AAC)

The 11172-3 MPEG-1 and IS13818-3 MPEG-2 BC/LSF algorithms standardized practical methods for high-quality coding of monaural and stereophonic program material. By the early 1990's, however, demand for high-quality coding of multichannel audio at reduced bit rates had increased significantly. Although the MPEG-1 and MPEG-2 BC/LSF algorithms had exploited many of the audio coding research advances that had occurred since the late 1980's, a few recent tools still had not been adopted in the international standards. Moreover, the backward compatibility constraints imposed on the MPEG-2 BC/LSF algorithm made it impractical to code five-channel program material at rates below 640 kb/s. As a result, MPEG began standardization activities for a non-backward compatible advanced coding system targeting "indistinguishable" quality [264] at a rate of 384 kb/s for five full bandwidth channels. In less than three years, this effort led to the adoption of the IS13818-7 MPEG-2 Non-backward Compatible/Advanced Audio Coding (NBC/AAC) algorithm [265], a system that exceeded design goals and produced the desired quality at only 320 kb/s for five full bandwidth channels. While similar in many respects to its predecessors, the AAC algorithm [75], [266], [267] achieves performance improvements by incorporating coding tools previously not found in the standards such as filter bank window shape adaptation, spectral coefficient prediction, temporal noise shaping, and bandwidth- and bit-rate-scaleable operation. Improvements in bit rate and quality are also realized through the use of a sophisticated noiseless coding scheme integrated with a two-stage bit allocation procedure. Moreover, the AAC algorithm contains scalability and complexity management tools not previously included with the MPEG algorithms. As far as applications are concerned, the AAC algorithm is currently embedded in the atob and LiquidAudio players for streaming of high-fidelity stereophonic audio. It is also a candidate for standardization in the United States Digital Audio Radio (U.S. DAR) project. The remainder of this section describes some of the features unique to MPEG-2 AAC.

The MPEG-2 AAC algorithm (Fig. 41) is organized as a set of coding tools. Depending upon available CPU or channel resources and desired quality, one can select from among three complexity "profiles," namely main, low (LC), and scalable sample rate (SSR) profiles. Each profile recommends a specific combination of tools. Our focus here is on the complete set of tools available for main profile coding, which works as follows.

1) Filter Bank: First, a high-resolution MDCT filter bank obtains a spectral representation of the input. Like previous MPEG coders, the AAC filter bank resolution is signal adaptive. Stationary signals are analyzed with a 2048-point window, while transients are analyzed with a block of eight 256-point windows to maintain time synchronization for channels using different filter bank resolutions during multichannel operations. The maximum frequency resolution is therefore 23 Hz for a 48 kHz sample rate, and the maximum time resolution is 2.6 ms. Unlike previous MPEG coders, however, AAC eliminates the hybrid filter bank and relies on the MDCT exclusively. The AAC filter bank is also unique in its ability to switch between two distinct MDCT analysis window shapes. Given particular input signal characteristics, the idea behind window shape adaptation is to optimize filter bank frequency selectivity in the sense of localizing supramasking threshold signal energy to the extent possible in the fewest spectral coefficients. This strategy seeks essentially to maximize the perceptual coding gain of the filter bank. While both satisfying the perfect reconstruction and aliasing cancellation constraints of the MDCT, the two windows offer different spectral analysis properties. A sine window [(47)] is selected when

narrow passband selectivity is more beneficial than strong stopband attenuation, as in the case of inputs characterized by a dense harmonic structure (less than 140-Hz spacing) such as harpsichord or pitch pipe. On the other hand, a KBD window is selected in cases for which stronger stopband attenuation is required, or for situations in which strong components are separated by more than 220 Hz. The KBD window in AAC has its origins in the MDCT filter bank window designed at Dolby Labs for the AC-3 algorithm using explicitly perceptual criteria. Details of the minimum masking template design procedure are given in [268] and [269].

2) Spectral Prediction: The AAC algorithm realizes improved coding efficiency relative to its predecessors by applying prediction over time to the transform coefficients below 16 kHz, as was done previously in [118], [270], and [271].

3) Bit Allocation: The bit allocation and quantization strategies in AAC bear some similarities to previous MPEG coders in that they make use of a nested loop iterative procedure, and in that psychoacoustic masking thresholds are obtained from an analysis model similar to MPEG-1, model recommendation number two. Both lossy and lossless coding blocks are integrated into the rate-control loop structure so that redundancy removal and irrelevancy reduction are simultaneously affected in a single analysis-by-synthesis process. As in the case of MPEG-1, Layer III, the AAC coefficients are grouped into 49 scale-factor bands that mimic the auditory system's frequency resolution. As with MPEG-1 Layer III and Lucent Technologies PAC, a bit reservoir is maintained to compensate for time-varying perceptual bit-rate requirements.

4) Noiseless Coding: The noiseless coding block [272] embedded in the rate-control loop has several innovative features as well. Twelve Huffman codebooks are available for two- and four-tuple blocks of quantized coefficients. Sectioning and merging techniques are applied to maximize redundancy reduction. Individual codebooks are applied to time-varying "sections" of scale-factor bands, and the sections are defined on each frame through a greedy merge algorithm that minimizes the bitrate. Grouping across time and intraframe frequency interleaving of coefficients prior to codebook application are also applied to maximize zero coefficient runs and further reduce bit rates.

5) Other Enhancements: The AAC has an embedded TNS module for pre-echo control (Section III-E), a special profile for SSR, and time-varying as well as frequency subband selective application of MS and/or intensity stereo coding for five-channel inputs [273].

6) Performance: Incorporation of the nonbackward compatible coding enhancements proved to be a judicious strategy for the AAC algorithm. In independent listening tests conducted worldwide [274], the AAC algorithm met the strict ITU-R BS.1116 criteria for "indistinguishable" quality [275] at a rate of 320 kb/s for five full bandwidth channels [276]. This level of quality was achieved with a manageable decoder complexity. Two-channel real-time AAC decoders were reported to run on 133-MHz Pentium platforms using 40% and 25% of available CPU resources for the main and low complexity profiles, respectively [277]. In the future, MPEG-2 AAC will maintain a presence as the core "time-frequency" coder reference model for the new MPEG-4 standard.

7) Reference Model Validation : Before proceeding with a discussion of MPEG-4, we first consider a significant system-level aspect of MPEG-2 AAC that also propagated into MPEG-4. Both algorithms are structured in terms of so-called reference models (RM's). In the RM approach, generic coder blocks or tools (e.g., perceptual model, filter bank, rate-control loop, etc.) adhere to a set of defined interfaces. The RM therefore facilitates the testing of incremental single block improvements without disturbing the existing macroscopic RM structure. For instance, one could devise a new psychoacoustic analysis model that satisfies the AAC RM interface and then simply replace the existing RM perceptual model in the reference software with the proposed model. It is then a straightforward matter to construct performance comparisons between the RM method and the proposed method in terms of quality, complexity, bit rate, delay, or robustness. The RM definitions are intended to expedite the process of evolutionary coder improvements.

In fact, several practical AAC improvements have already been analyzed within the RM framework. For example, in [278] a new backward predictor is proposed as a replacement for the existing backward adaptive LMS predictors, resulting in a 38% computational savings. Forward adaptive predictors have also been investigated [279]. In another example of RM efficacy, improvements to the AAC noiseless coding module were also reported in [280]. A modification to the greedy merge sectioning algorithm was proposed in which high-magnitude spectral peaks that tended to degrade Huffman coding efficiency were coded separately. In yet another example of RM innovation aimed at improving quality for a given bit rate, product code VQ techniques [281] were applied to increase AAC scale-factor coding efficiency [282]. This scheme realized significant quality improvements for critical test items at low rates, because scale factors are decorrelated using a DCT and then grouped into subvectors for quantization by a product code VQ [283].

8) Enhanced AAC in MPEG-4: The next section is concerned with the multimodal MPEG-4 audio standard, for which the MPEG-2 AAC RM core was selected as the "time-frequency" audio coding RM, although some improvements have already been realized. Recently, for example, perceptual noise substitution (PNS) was included [284] as part of the MPEG-4 AAC RM. The PNS exploits the fact that a random noise process can be used to model efficiently transform coefficients in noise-like frequency subbands, provided the noise vector has an appropriate temporal fine structure [122]. Bit-rate reduction is realized since only a compact, parametric representation is required for each PNS subband (i.e., noise energy) rather than requiring full quantization and coding of subband transform coefficients. At a bit rate of 32 kb/s, a mean improvement due to PNS of $+0.61$ on the comparison mean opinion score

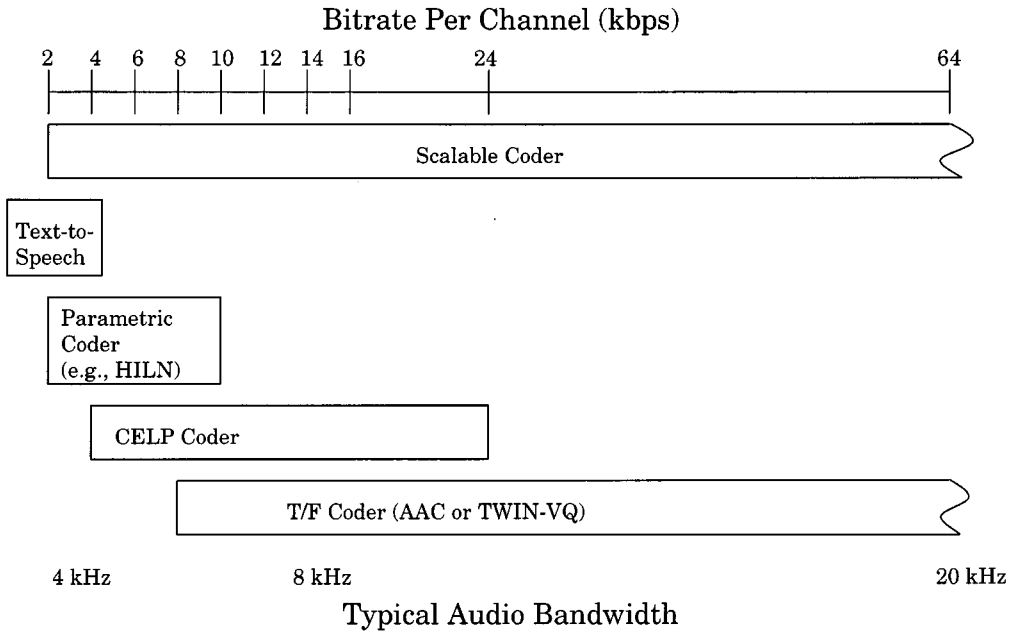


Fig. 42. ISO/IEC MPEG-4 integrated tools for audio coding (after [288]).

(CMOS) test for critical test items such as speech, castanets, and complex sound mixtures was reported in [284].

C. ISO/IEC 14 496-3 (MPEG-4)

Version one of the most recent MPEG audio standard, ISO/IEC 14 496 or MPEG-4, was adopted in December 1998 after many proposed algorithms were tested [285], [286] for compliance with the program objectives established by the MPEG committee. MPEG-4 audio encompasses a great deal more functionality than just perceptual coding [287]. It comprises an integrated family of algorithms with wide-ranging provisions for scaleable, object-based speech and audio coding at bit rates from as low as 200 b/s up to 64 kb/s per channel. The distinguishing features of MPEG-4 relative to its predecessors are extensive scalability, object-based representations, user interactivity/object manipulation, and a comprehensive set of coding tools available to accommodate almost any desired tradeoff among bit rate, complexity, and quality. Very low rates are achieved through the use of structured representations for synthetic speech and music, such as text-to-speech and MIDI. For higher bit rates and "natural audio" speech and music, the standard provides integrated coding tools that make use of different signal models, the choice of which is made depending upon desired bit rate, bandwidth, complexity, and quality. Coding tools are also specified in terms of MPEG-4 "profiles," which essentially recommend tool sets for a given level of functionality and complexity. Beyond its provisions specific to coding of speech and audio, MPEG-4 also specifies numerous sophisticated system-level functions for media-independent transport, efficient buffer management, syntactic bitstream descriptions, and time-stamping for synchronization of audiovisual information units. Although a discussion of these features is not relevant to our focus on perceptual coding, an excellent overview is given in [288]. Also note that a perspective on future directions within MPEG audio appeared in [289].

1) Natural Audio Coding Tools: MPEG-4 audio version one [288] integrates a set of tools (Fig. 42) for coding of natural sounds [290] at bit rates ranging from as low as 200 b/s up to 64 kb/s per channel. For speech and audio, three distinct algorithms are integrated into the framework, namely, two parametric coders for bitrates of 2–4 kb/s and 8-kHz sample rate as well as 4–16 kb/s and 8- or 16-kHz sample rates (Section VI-B). For higher quality, narrow-band (8-kHz sample rate) or wide-band (16 kHz) speech is handled by a CELP speech codec operating between 6 and 24 kb/s. For generic audio at bit rates above 16 kb/s, a "time/frequency" perceptual coder is employed, and in particular the MPEG-2 AAC algorithm with extensions for fine-grain bit-rate scalability [291] is specified in MPEG-4 version one RM as the time-frequency coder. The multimodal framework of MPEG-4 audio allows the user to tailor the coder characteristics (i.e., the signal model) to the program material.

2) Synthetic Audio Coding Tools: Whereas earlier MPEG standards treated only natural audio program material, MPEG-4 achieves very low rate coding by supplementing its natural audio coding techniques with tools for synthetic audio processing [292] and interfaces for structured, high-level audio representations. Chief among these are the text-to-speech interface (TTSI) and methods for score-driven synthesis. The TTSI provides the capability for 200–1200 b/s transmission of synthetic speech that can be represented in terms of either text only or text plus prosodic parameters. Beyond speech, general music synthesis capabilities in MPEG-4 are provided by a set of structured audio tools [293]–[295]. Synthetic sounds are represented using the structured audio orchestra language (SAOL). SAOL

[296] treats music as a collection of instruments and instruments as small networks of signal-processing primitives, all of which can be downloaded to a decoder. Although no standard synthesis techniques are specified, available synthesis methods include the following: wavetable, FM, additive, physical modeling, granular synthesis, or nonparametric hybrids of any of these methods [297]. An excellent tutorial on structured audio methods and applications appeared recently in [298].

3) MPEG-4 Audio Profiles: Although many coding and processing tools are available in MPEG-4 audio, cost and complexity constraints often dictate that it is not practical to implement all of them in a particular system. Version 1 therefore defines four complexity-ranked audio profiles intended to help system designers in the task of appropriate tool subset selection. In order of bit rate, they are as follows. The low-rate synthesis audio profile provides only wavetable-based synthesis and a text-to-speech (TTS) interface. For natural audio-processing capabilities, the speech audio profile provides a very low-rate speech coder and a CELP speech coder. The scaleable audio profile offers a superset of the first two profiles. With bit rates ranging from 6 to 24 kb/s and bandwidths from 3.5 to 9 kHz, this profile is suitable for scalable coding of speech, music, and synthetic music in applications such as Internet streaming or narrow-band audio digital broadcasting (NADIB). Finally, the main audio profile is a superset of all other profiles, and it contains tools for both natural and synthetic audio.

4) MPEG-4 Audio Version Two: While remaining backward compatible with MPEG-4 version 1, MPEG-4 version 2 will add new profiles that incorporate a number of significant system-level and functionality enhancements. At the system level, version 2 will include a media independent bitstream format that supports streaming, editing, local playback, and interchange of contents. Also in version 2, an MPEG-J "programmatic system" will specify an application programming interface (API) for interoperation of MPEG players with JAVA code. New error resilience techniques in version 2 will allow both equal and unequal error protection for the audio bit streams. As for functionality, version 2 will offer improved audio realism in sound rendering. New tools will allow parameterization of the acoustical properties of an audio scene, enabling features such as immersive audio-visual rendering, room acoustical modeling, and enhanced three-dimensional sound presentation. TTS interfaces from version 1 will be enhanced in version 2 with a markup TTS intended for applications such as speech-enhanced Web browsing, verbal e-mail, and "story-teller" on demand. MPEG-4 standardization activities are ongoing. One can obtain up-to-date information from several on-line sources. For example, structured audio information can be found on [299]. The complete 2500 page May 1998 MPEG-4 Final Committee Draft document is also available electronically from [299].

D. Precision Adaptive Subband Coding

Phillips' DCC is an example of a consumer product that 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 [300].

E. Adaptive Transform Acoustic Coding

The ATRAC algorithm developed by Sony for use in its rewritable MiniDisc system makes combined use of subband and transform coding techniques to achieve nearly CD-quality coding of 44.1-kHz 16-bit PCM input data [301] at a bit rate of 146 kb/s per channel [302]. Using a tree-structured QMF analysis bank, the ATRAC encoder (Fig. 43) first splits the input signal into three subbands of 0–5.5 kHz, 5.5–11 kHz, and 11–22 kHz. Like MPEG layer III, the ATRAC QMF bank is followed by signal-adaptive MDCT analysis [(44)] in each subband. The window switching scheme works as follows. During steady-state input periods, high-resolution spectral analysis is attained using 512 sample blocks (11.6 ms). During input attack or transient periods, however, 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. After MDCT analysis, spectral components are clustered into 52 nonuniform subbands [block floating units (BFU's)] according to a critical band spacing. The BFU's are block-companded, quantized, and encoded according to a psychoacoustically derived bit allocation. For each analysis frame, the ATRAC encoder transmits quantized MDCT coefficients, subband window lengths, BFU scalefactors, and BFU word lengths to the decoder. Like the MPEG family, the ATRAC architecture decouples the decoder from psychoacoustic analysis and bit allocation details. Evolutionary improvements in the encoder bit allocation strategy are therefore possible without modifying the decoder structure. An added benefit of this architecture is asymmetric complexity, which enables inexpensive decoder implementations.

Suggested bit allocation techniques for ATRAC are of lower complexity than those found in other standards since ATRAC is intended for low-cost, battery-powered consumer electronics equipment. One proposed method distributes bits between BFU's according to a weighted combination of fixed and adaptive bit allocations [303]. For the kth BFU, bits are allocated according to the relation

$$r(k) = \alpha \cdot r_a(k) + (1-\alpha) \cdot r_f(k) - \beta \qquad (52)$$

where

$r_f(k)$	fixed allocation;
$r_a(k)$	signal-adaptive allocation;
parameter β	constant offset computed to guarantee a fixed bit rate;
parameter α	tonality estimate ranging from zero (noise-like) to one (tone-like).

The fixed allocations $r_f(k)$ are the same for all inputs and concentrate more bits at lower frequencies. The signal-adaptive bit allocations $r_a(k)$ allocate bits according to the strength of the MDCT components. The effect of (52) is that more bits are allocated to BFU's containing strong peaks for tonal signals. For noise-like signals, bits are allocated

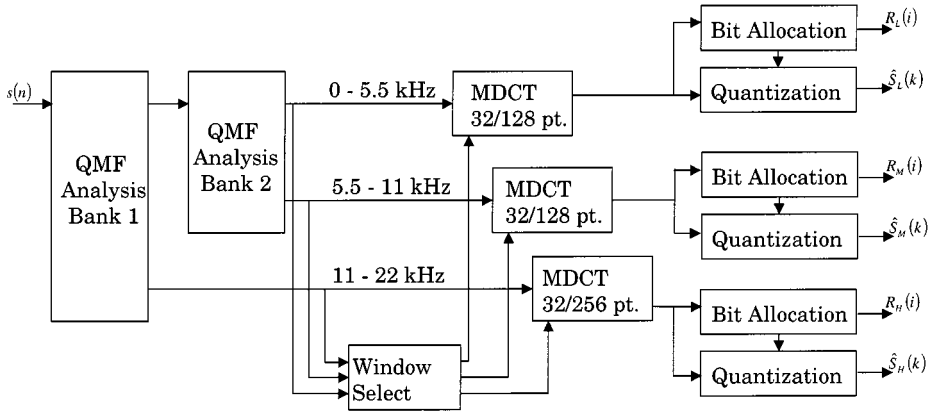


Fig. 43. Sony ATRAC (MiniDisc, SDDS).

according to the fixed allocation, with low bands receiving more bits than high bands. Clearly, this method relies on heuristic principles rather than detailed psychoacoustic analysis such as the MPEG model recommendations (Section VIII-G). The resulting system achieves a reasonable tradeoff among complexity, quality, and bit rate.

F. Sony Dynamic Digital Sound (SDDS)

In addition to enabling near CD quality on a MiniDisc medium, the ATRAC algorithm has also been deployed as the core of Sony's digital cinematic sound system, SDDS. SDDS integrates eight independent ATRAC modules to carry the program information for the left, left center, center, right center, right, subwoofer, left surround, and right surround channels typically present in a modern theater. SDDS data are recorded using optical black and white dot-matrix technology onto two thin strips along the right and left edges of the film, outside of the sprocket holes, and each edge contains four channels. There are 512 ATRAC bits per channel associated with each movie frame, and each optical data frame contains a matrix of 52×192 bits [304]. SDDS data tracks do not interfere with or replace the existing analog sound tracks. Both Reed–Solomon error correction and redundant track information delayed by 18 frames are employed to make SDDS robust to bit errors introduced by run-length scratches, dust, splice points, and defocusing during playback or film printing. Analog program information is used as a backup in the event of uncorrectable digital errors.

G. Lucent Technologies Perceptual Audio Coder (PAC), Enhanced PAC (EPAC), and Multichannel PAC (MPAC)

The pioneering research contributions on perceptual entropy [45], monophonic PXFM [6], stereophonic PXFM [305], and ASPEC [9] strongly influenced not only the MPEG family architecture but also evolved at AT&T Bell Laboratories into the PAC. AT&T and Lucent Technologies separated after the MPAC algorithm was evaluated for MPEG NBC/AAC testing, and the PAC algorithm subsequently became proprietary to Lucent. AT&T, meanwhile, has become active in the MPEG-2 AAC research and standardization. The low-complexity profile of AAC has become the AT&T coding standard. Like the MPEG coders, the current Lucent PAC algorithm is flexible in that it supports monophonic, stereophonic, and multiple channel modes. In fact, the bitstream definition will accommodate up to 16 front side, seven surround, and seven auxiliary channel pairs, as well as three low-frequency effects (LFE or subwoofer) channels. Depending upon desired quality, PAC supports several bit rates. For a modest increase in complexity at a particular bit rate, moreover, improved output quality can be realized by enabling enhancements to the original system (EPAC). For example, whereas 96-kb/s output was judged to be adequate with stereophonic PAC, near and transparent CD output qualities were reported at 56–64 kb/s and 128 kb/s, respectively, for stereophonic EPAC [306]. This section gives an overview of the PAC, EPAC, and MPAC algorithms, concentrating primarily on the innovations that differentiate this system from the others reviewed in this document.

1) PAC: The original PAC system described in [307] achieves very high-quality coding of stereophonic inputs at 96 kb/s. Like MPEG-1 layer III and ATRAC, the PAC encoder [Fig. 44(a)] uses a signal-adaptive MDCT filter bank to analyze the input spectrum with appropriate frequency resolution. A long window of 2048 points (1024 subbands) is used during steady-state segments, or else a series of short 256-point windows (128 subbands) is applied during segments containing transients or sharp attacks. In contrast to MPEG-1 and ATRAC, however, PAC relies on the MDCT alone rather than incorporating MDCT analysis into a hybrid filter bank structure, thus realizing a relative complexity reduction in the filter bank section. As noted previously [115], [119], the MDCT lends itself to compact representation of stationary signals, and a 2048-point block size yields sufficiently high frequency resolution for most sources. This segment length was also associated with the maximum realizable coding gain as a function of block size [308]. Masking thresholds are used to select one of 128 exponentially distributed quantization step sizes in each of 49 or 14 coder bands (analogous to ATRAC BFU's) in high-resolution and low-resolution modes, respectively. The coder bands are quantized using an iterative rate control loop in which thresholds are adjusted to satisfy simultaneously

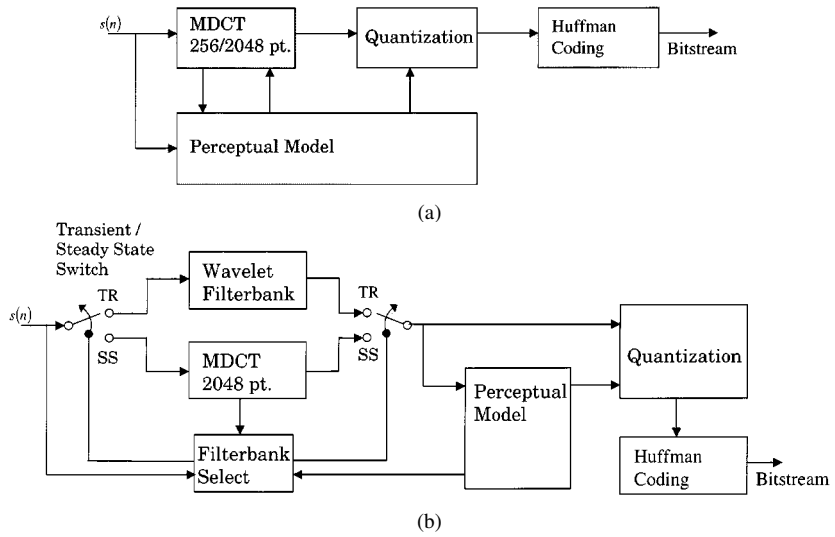


Fig. 44. Lucent Technologies PAC: (a) PAC and (b) EPAC.

bit-rate constraints and an equal loudness criterion that attempts to shape quantization noise such that its absolute loudness is constant relative to the masking threshold. The rate control loop allows time-varying instantaneous bit rates, much like the bit reservoir of MPEG-1 layer III. Remaining statistical redundancies are removed from the stream of quantized spectral samples prior to bitstream formatting using eight structured, multidimensional Huffman codebooks.

2) EPAC: In an effort to enhance PAC output quality at low bitrates, Sinha and Johnston introduced a novel signal-adaptive MDCT/WP switched filter bank scheme [Fig. 44(b)], which resulted in nearly transparent coding for CD-quality source material at 64 kb/s per stereo pair [308]. EPAC is unique in that it switches between two distinct filter banks rather than relying upon hybrid [17], [302] or nonuniform cascade [201] structures. In subjective tests involving 12 expert and nonexpert listeners with difficult castanets and triangle test signals, EPAC outperformed PAC at a rate of 64-kb/s per stereo pair by an average of 0.4–0.6 on a five-point quality scale.

3) MPAC: Like the MPEG, AC-3, and SDDS systems, the PAC algorithm also extends its monophonic processing capabilities into stereophonic and multiple channel modes. Stereophonic PAC computes individual masking thresholds for the left, right, mono, and stereo (L, R, $M = L + R$, and $S = L - R$) signals using a version of the monophonic perceptual model that has been modified to account for binary-level masking differences (BLMD's) or binaural unmasking effects [309]. Then, monaural PAC methods encode either the signal pairs L, R or M, S. In order to minimize the overall bit rate, however, an LR/MS switching procedure is embedded in the rate control loop such that different encoding modes (LR or MS) can be applied to the individual coder bands on the same analysis frame. MPAC was found to produce the best quality at 320 kb/s for five channels during a recent ISO test of multichannel algorithms [310].

4) Applications: Both 128- and 160-kb/s stereophonic versions of PAC are currently being considered for standardization in the U.S. DAR project. In an effort to provide graceful degradation and extend broadcast range in the presence of heavy fading associated with fringe reception areas, perceptually motivated unequal error protection (UEP channel coding) schemes were examined in [311]. The availability of JAVA PAC decoder implementations are reportedly increasing PAC deployment among suppliers of internet audio program material [306]. MPAC has been considered for cinematic and advanced television applications. Real-time PAC and EPAC decoder implementations have been demonstrated on 486-class PC platforms.

H. DOLBY AC-2, AC-2A

Since the late 1980's, Dolby Laboratories has been active in perceptual audio coding research and standardization, and Dolby researchers have made numerous scientific contributions within the collaborative framework of MPEG audio. On the commercial front, Dolby has developed the AC-2 and the AC-3 algorithms [268]. The AC-2 [312], [313] is a family of single-channel algorithms operating at bit rates between 128 and 192 kb/s for 20-kHz bandwidth input sampled at 44.1 or 48 kHz. There are four available AC-2 variants, all of which share a common architecture in which the input is mapped to the frequency domain by an evenly stacked TDAC filter bank [87] with a novel parametric Kaiser–Bessel analysis window (Sections III-C and VIII-B) optimized for improved stopband attenuation relative to the sine window. The evenly stacked TDAC differs from the oddly stacked MDCT in that the evenly stacked low-band filter is half-band, and its magnitude response wraps around the foldover frequency (see Section III). A unique mantissa-exponent coding scheme is applied to the TDAC transform coefficients. First, sets of frequency-adjacent coefficients are grouped into blocks (subbands) of roughly critical bandwidth. For each, the block maximum is identified and then quantized as an exponent in terms of the number of left shifts required until overflow occurs. The collection of exponents forms a stair-step spectral envelope having 6 dB (left shift =

multiply by $2 = 6.02$ dB) resolution, and normalizing the transform coefficients by the envelope generates a set of mantissas. The envelope approximates the short-time spectrum, and therefore a perceptual model uses the exponents to compute both a fixed and a signal-adaptive bit allocation for the mantissas on each frame. As far as details on the four AC-2 variants are concerned, two versions are designed for low-complexity, low-delay applications, and the other two for higher quality at the expense of increased delay or complexity. The AC-2A [314] algorithm employs a switched 128/512-point TDAC filter bank to improve quality for transient signals. One AC-2 feature that is unique among the standards is that the perceptual model is backward adaptive, meaning that the bit allocation is not transmitted explicitly. Instead, the AC-2 decoder extracts the bit allocation from the quantized spectral envelope using the same perceptual model as the AC-2 encoder. This structure leads to a significant reduction of side information and induces a symmetric encoder/decoder complexity, which was well suited to the original AC-2 target application of single point-to-point audio transport. An example single point-to-point system now using low-delay AC-2 is the DolbyFAX, a full-duplex codec that carries simultaneously two channels in both directions over four ISDN "B" links for film and TV studio distance collaboration. Low-delay AC-2 codecs have also been installed on 950 MHz wireless digital studio transmitter links (DSTL's). The AC-2 moderate delay and AC-2A algorithms have been used for both network and wireless broadcast applications such as cable and DBS television.

I. Dolby AC-3/Dolby Digital/Dolby SR·D

The 5.1-channel "surround" format that had become the *de facto* standard in most movie houses during the 1980's was becoming ubiquitous in home theaters of the 1990's that were equipped with matrixed multichannel sound (e.g., Dolby ProLogic). As a result of this trend, it was clear that emerging applications for perceptual coding would eventually minimally require stereophonic or even multichannel surround-sound capabilities to gain consumer acceptance. Although single-channel algorithms such as the AC-2 can run on parallel independent channels, significantly better performance can be realized by treating multiple channels together in order to exploit interchannel redundancies and irrelevancies. The Dolby Laboratories AC-3 algorithm [315]–[317], also known as "Dolby Digital" or "SR·D," was developed specifically for multichannel coding by refining all of the fundamental AC-2 blocks, including the filter bank, the spectral envelope encoding, the perceptual model, and the bit allocation. The coder carries 5.1 channels of audio (left, center, right, left surround, right surround, and a subwoofer), but at the same time it incorporates a flexible downmix strategy at the decoder to maintain compatibility with conventional monaural and stereophonic sound reproduction systems. The ".1" channel is usually reserved for low-frequency effects and is low-pass bandlimited below 120 Hz. The main features of the AC-3 algorithm are as follows:

- sample rates: 32, 44.1, and 48 kHz;
- high-quality output at 64 kb/s per channel;
- MDCT filter bank (TDAC [90]), KBD window;
- spectral envelope represented by exponents;
- hybrid forward–backward adaptive perceptual model;
- uniform quantization of mantissas;
- multiple channels processed as an ensemble;
- robust decoder downmix functionality;
- board-level real-time encoders available;
- bit rates: 32–640 kb/s, variable;
- delay roughly 100 ms;
- exponents/mantissa quantization/encoding;
- signal-adaptive exponent strategy;
- parametric bit allocation;
- perceptual model improvements possible;
- frequency-selective intensity coding, LR, MS;
- integral dynamic range control system;
- chip-level real-time decoders available.

The AC-3 works in the following way. A signal-adaptive MDCT filter bank with a customized KBD window (Sections III-C and VIII-B) maps the input to the frequency domain. Long windows are applied during steady-state segments, and a pair of short windows is used for transient segments. The MDCT coefficients are quantized and encoded by an exponent/mantissa scheme similar to AC-2. Bit allocation for the mantissas is performed according to a perceptual model that estimates the masked threshold from the quantized spectral envelope. Like AC-2, an identical perceptual model resides at both the encoder and decoder to allow for backward adaptive bit allocation on the basis of the spectral envelope, thus reducing the burden of side information on the bitstream. Unlike AC-2, however, the perceptual model is also forward adaptive in the sense that it is parametric. Model parameters can be changed at the encoder and the new parameters transmitted to the decoder in order to affect modified masked threshold calculations. Particularly at lower bit rates, the perceptual bit allocation may yield insufficient bits to satisfy both the masked threshold and the rate constraint. When this happens, mid/side (MS) and intensity coding ("channel coupling" above 2 kHz) reduce the demand for bits by exploiting, respectively, interchannel redundancies and irrelevancies. Ultimately, exponents, mantissas, coupling data, and exponent strategy data are combined and transmitted to the receiver.

1) Filter Bank: Although the high-level AC-3 structure (Fig. 45) resembles that of AC-2, there are significant differences between the two algorithms. Like AC-2, the AC-3 algorithm first maps input samples to the frequency domain using a PR cosine-modulated filter bank with a novel KBD window (Sections III-C and VIII-B, parameters in [268]). Unlike AC-2, however, AC-3 is based on the oddly stacked MDCT. The AC-3 also handles window switching differently than AC-2A. Long, 512-sample (93.75 Hz res. at 48 kHz) windows are used to achieve reasonable coding gain during stationary segments. During transients, however, a pair of 256-sample windows replaces the long window to minimize pre-echoes. Also in contrast to the MPEG and AC-2 algorithms, the AC-3 MDCT filter bank retains PR properties

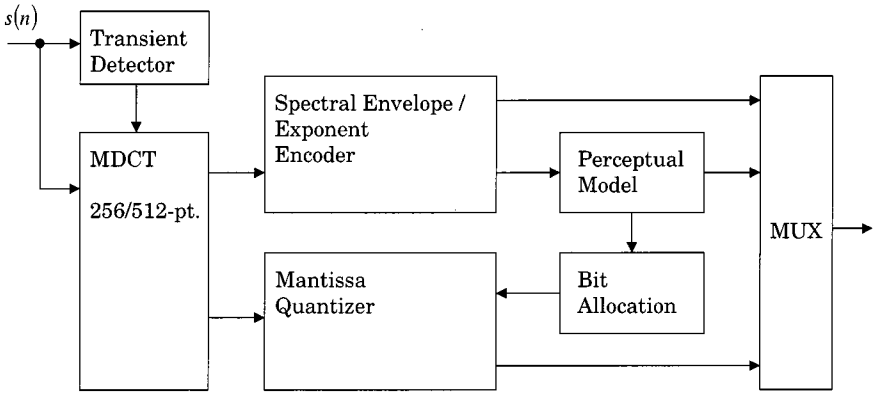


Fig. 45. Dolby AC-3 encoder.

during window switching without resorting to bridge windows by introducing a suitable phase shift into the MDCT basis vectors (equations given in [106]) for one of the two short transforms. Whenever a scheme similar to the one used in AC-2A detects transients, short filter bank windows may activate independently on any one or more of the 5.1 channels.

2) Exponent Strategy: The AC-3 algorithm uses a refined version of the AC-2 exponent/mantissa MDCT coefficient representation, resulting in a significantly improved coding gain. In AC-3, the MDCT coefficients corresponding to 1536 input samples (six transform blocks) are combined into a single frame. Then, a frame processing routine optimizes the exponent representation to exploit temporal redundancy, while at the same time representing the stair-step spectral envelope with adequate frequency resolution. In particular, spectral envelopes are formed from partitions of either one, two, or four consecutive MDCT coefficients on each of the six MDCT blocks in the frame. To exploit time redundancy, the six envelopes can be represented individually, or any or all of the six can be combined into temporal partitions. The AC-3 exponent strategy exploits in a signal-dependent fashion the time- and frequency-domain redundancies that exist on a frame of MDCT coefficients.

3) Perceptual Model: A novel parametric forward–backward adaptive perceptual model estimates the masked threshold on each frame. The forward-adaptive component exists only at the encoder. Given a rate constraint, this block interacts with an iterative rate control loop to determine the best set of perceptual model parameters. These parameters are passed to the backward adaptive component, which estimates the masked threshold by applying the parameters from the forward-adaptive component to a calculation involving the quantized spectral envelope. Identical backward adaptive model components are embedded in both the encoder and decoder. Thus, model parameters are fixed at the encoder after several threshold calculations in an iterative rate control process and then transmitted to the decoder. The parametric perceptual model also provides a convenient upgrade path in the form of a bit allocation delta parameter. It was envisioned that future, more sophisticated AC-3 encoders might run in parallel two perceptual models, with one being the original reference model and the other being an enhanced model with more accurate estimates of masked threshold. The delta parameter allows the encoder to transmit a stair-step function for which each tread specifies a masking level adjustment for an integral number of 1/2-Bark bands. Thus, the masking model can be incrementally improved without alterations to the existing decoders. Other details on the hybrid backward-forward AC-3 perceptual model can be found in [269].

4) Bit Allocation and Mantissa Quantization: A bit allocation is determined at the encoder for each frame of mantissas by an iterative procedure that adjusts the mantissa quantizers, the multichannel coding strategies (below), and the forward-adaptive model parameters to satisfy simultaneously the specified rate constraint and the masked threshold. In a manner similar to MPEG-1, quantizers are selected for the set of mantissas in each partition based on an SMR calculation. Sufficient bits are allocated to ensure that the SNR for the quantized mantissas is greater than or equal to the SMR. If the bit supply is insufficient to satisfy the masked threshold, then SNR's can be reduced in selected threshold partitions until the rate is satisfied, or intensity coding and MS transformations are used in a frequency-selective fashion to reduce the bit demand. Unlike some of the other standardized algorithms, the AC-3 does not include an explicit lossless coding stage for final redundancy reduction after quantization and encoding.

5) Multichannel Coding: When bit demand imposed by multiple independent channels exceeds the bit budget, the AC-3 ensemble processing of 5.1 channels exploits interchannel redundancies and irrelevancies, respectively, by making frequency-selective use of MS and intensity coding techniques. Although the MS and intensity functions can be simultaneously active on a given channel, they are restricted to nonoverlapping subbands. The MS scheme is carefully controlled [317] to maintain compatibility between AC-3 and matrixed surround systems such as Dolby ProLogic. Intensity coding, also known as channel coupling, is a multichannel irrelevancy reduction coding technique that exploits properties of spatial hearing. There is considerable experimental evidence [318] suggesting that the interaural time difference of a signal's fine structure has negligible influence on sound localization above a certain frequency. Instead, the ear evaluates primarily energy envelopes. Thus,

the idea behind intensity coding is to transmit only one envelope in place of two or more sufficiently correlated spectra from independent channels, together with some side information. The side information consists of a set of coefficients that is used to recover individual spectra from the intensity channel.

6) System-Level Functions: At the system level, AC-3 provides mechanisms for channel downmixing and dynamic range control. Downmix capability is essential for the 5.1 channel system since the majority of potential playback systems are still monaural or, at best, stereophonic. Downmixing is performed at the decoder in the frequency domain rather than the time domain to minimize complexity. This is possible because of the filter bank linearity. The bitstream carries some downmix information since different listening situations call for different downmix weighting. Dialog level normalization is also available at the decoder. Finally, the bitstream has available facilities to handle other control and ancillary user information such as copyright, language, production, and time-code data [319].

7) Complexity: Assuming the standard HDTV configuration of 384 kb/s with a 48-kHz sample rate and implementation using the Zoran ZR38001 general-purpose DSP instruction set, the AC-3 decoder memory requirements and complexity are as follows: 6.6 kb RAM, 5.4 kb ROM, 27.3 MIPS for 5.1 channels; and 3.1 kb RAM, 5.4 kb ROM, and 26.5 MIPS for two channels [320]. Note that complexity estimates are processor-dependent. For example, on a Motorola DSP56002, 45 MIPS are required for a 5.1-channel decoder. Encoder complexity varies between two and five times decoder complexity depending on the encoder sophistication [320]. Numerous real-time encoder and decoder implementations have been reported. Early on, for example, a single-chip decoder was implemented on a Zoran DSP [321]. More recently, a DP561 AC-3 encoder (5.1 channels, 44.1- or 48-kHz sample rate) for DVD mastering was implemented in real time on a DOS/Windows PC host with a plug-in DSP subsystem. The computational requirements were handled by an Ariel PC-Hydra DSP array of eight Texas Instruments TMS 320C44 floating point DSP devices clocked at 50 MHz [322]. The authors also reported on anticipated completion of a similar real-time encoder with only two or three 80-MHz fixed-point Motorola 56 300 DSP devices [322].

8) Applications and Standardization: The first popular AC-3 application was in the cinema. The "Dolby Digital" or "SR D" AC-3 information is interleaved between sprocket holes on one side of the 35-mm film. The AC-3 was first deployed in only three theaters for the film *Star Trek VI* in 1991, after which the official rollout of Dolby SR D occurred in 1992 with *Batman Returns*. By 1997, more than 900 film soundtracks had been AC-3 encoded. Nowadays, the AC-3 algorithm is finding use in DVD, cable television, and DBS. Many high-fidelity amplifiers and receiver units now contain embedded AC-3 decoders and accept an AC-3 digital rather than an analog feed from external sources such as DVD. In addition, the DP504/524 version of the DolbyFAX system (Section VIII-H) has added AC-3 stereo and MPEG-1 Layer II to the original AC-2-based system. Film, television, and music studios use DolbyFAX over ISDN links for automatic dialog replacement, music collaboration, sound-effects delivery, and remote videotape audio playback. As far as standardization is concerned, the U.S. Advanced Television Systems Committee (ATSC) has adopted the AC-3 algorithm as the A/52 audio compression standard [362] and as the audio component of the A/52 DTV standard [323]. The U.S. Federal Communications Commission in December 1996 adopted the ATSC standard for DTV, including the AC-3 audio component. On the international standardization front, the Digital Audio-Visual Council (DAVIC) selected AC-3 and MPEG-1, layer II for the audio component of the DAVIC 1.2 specification [324]. Moreover, the Society of Cable and Telecommunications Engineers has considered AC-3 for standardization.

IX. Quality Measures for Perceptual Audio Coding

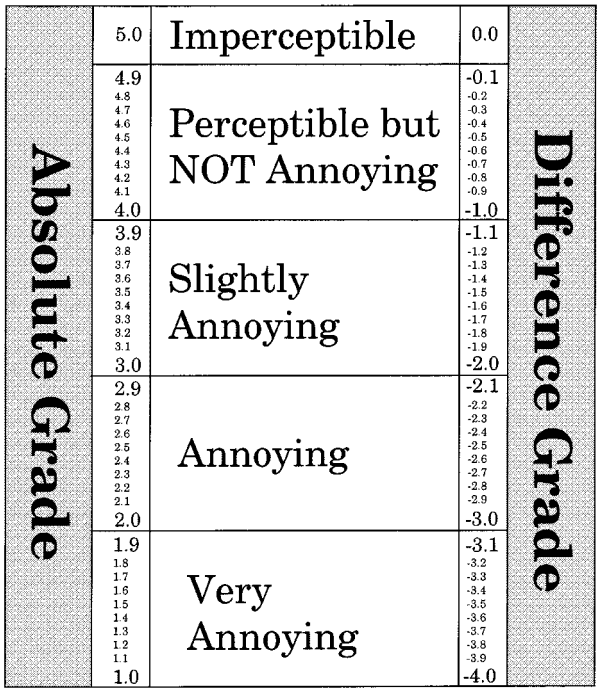
In many situations, and particularly in the context of standardization activities, performance measures are needed to evaluate whether one of the established or emerging techniques in perceptual audio coding is in some sense superior to the available alternative methods. Perceptual audio codecs are most often evaluated in terms of bit rate, complexity, delay, robustness, and output quality. Of these, all but robustness and output quality can be quantified in straightforward objective terms. Reliable and repeatable output quality assessment (which is related to robustness), on the other hand, presents a significant challenge. It is well known that perceptual coders can achieve transparent quality over a very broad, highly signal-dependent range of segmental SNR's ranging from as low as 13 dB to as high as 90 dB. Classical objective measures of signal fidelity such as SNR or total harmonic distortion (THD) are therefore completely inadequate [325]. As a result, time-consuming and expensive subjective listening tests are required to measure the small impairments that most often characterize the high-quality perceptual coding algorithms. Despite some confounding factors, subjective listening tests are nevertheless the most reliable tool available for codec quality evaluation, and standardized listening test procedures have been developed to maximize reliability. This section offers a perspective on quality measures for perceptual audio coding. The first portion describes subjective quality measurement techniques for perceptual audio coders and identifies confounding factors that complicate subjective tests, and the second portion gives sample subjective test results from several of the two- and 5.1-channel standards.

A. Subjective Quality Measures

Although listening tests are often conducted informally, the ITU-R Recommendation BS.1116 [275] formally specifies a listening environment and test procedure appropriate for subjective evaluations of the small impairments associated with high quality audio codecs. The standard procedure calls for grading by expert listeners [326] using the CCIR "continuous" impairment scale [Fig. 46(a)] [327] in a double blind, A-B-C triple-stimulus hidden reference

comparison paradigm. While stimulus A always contains the reference (uncoded) signal, the B and C stimuli contain in random order a repetition of the reference and then the impaired (coded) signal, i.e., either B or C is a hidden reference. After listening to all three, the subject must identify either B or C as the hidden reference, and then grade the impaired stimulus (coded signal) relative to the reference stimulus using the five-category, 41-point "continuous" absolute category rating (ACR) impairment scale shown in the left-hand column of Fig. 46(a). A default grade of 5.0 is assigned to the stimulus identified by the subject as the hidden reference. A subjective difference grade (SDG) is computed by subtracting the score assigned to the actual hidden reference from the score assigned to the actual impaired signal. Nearly transparent quality for the coded signal is implied if the hidden reference mean subjective score (MSS) lies within the 95% confidence interval of the coded signal and the coded signal MSS lies within the 95% confidence interval of the hidden reference. It is important to note the difference between the small impairment subjective measurements in [275] and the five-point MOS most often associated with speech coding algorithms [328]. Unlike the small impairment scale, the scale of the speech coding MOS is discrete, and scores are absolute rather than relative to a hidden reference. To emphasize this difference, it has been proposed [329] that MSS denote the small impairment subjective score for perceptual audio coders. Unless otherwise specified, the subjective listening test scores cited for the various algorithms described in this paper are from either the absolute or the differential small impairment scales in Fig. 46(a).

It is important to realize that the most reliable subjective evaluation strategy for a given perceptual codec depends on the nature of the coding distortion. Although the small-scale impairments associated with nearly transparent coding are well characterized by measurements relative to a reference standard using a fine-grade scale, some experts have argued that the more audible distortions associated with nontransparent coding are best measured using a different scale that can better cope with large impairments. For example, in recent listening tests [330] on 16-kb/s codecs for the WorldSpace satellite communications system, it was determined that an ITU-T P.800/P.830 seven-point comparison category rating (CCR) method [331] was better suited to the evaluation task [Fig. 46(b)] than the scale of BS.1116 because of the nontransparent quality associated with the test signal. Investigators preferred the CCR over both the small impairment scale as well as the five-point ACR commonly used in tests of speech codecs. A listening test standard for large-scale impairments analogous to BS.1116 does not yet exist for audio codec evaluation.



(a)

CCR	
A much better than B	+3
A better than B	+2
A slightly better than B	+1
A same as B	0
A slightly worse than B	-1
A worse than B	-2
A much worse than B	-3

(b)

Fig. 46. Subjective quality scales: (a) ITU-R Rec. BS.1116 [275] small impairment scale for absolute and differential subjective quality grades and (b) ITU-T Rec. P.800/P.830 [331] large impairment comparison category rating.

B. Confounding Factors in Subjective Evaluations

Regardless of the particular grading scale in use, subjective test outcomes generated using even rigorous methodologies such as the ITU-R BS.1116 are still influenced by factors such as context, site selection, and individual listener acuity (physical) or preference (cognitive). Before comparing subjective test results on particular codecs, therefore, one should be prepared to interpret the subjective scores with some care. For example, consider the variability of "expert" listeners. A study of decision strategies [332] using multidimensional scaling techniques [333] found that subjects disagree on the relative importance with which to weigh perceptual criteria during impairment detection tasks. In another study [334], Shlien and Soulodre presented experimental evidence that can be interpreted as a repudiation of the "golden ear." Expert listeners were tasked with discriminating between clean audio and audio corrupted by low-level artifacts typically induced by audio codecs (five types were analyzed in [335]), including pre-echo distortion, unmasked granular (quantization) noise, and high-frequency boost or attenuation. Different experts were sensitive to different artifact types. Sporer reached

similar conclusions after yet a third study of expert listeners [329]. Nonhuman factors also influence subjective listening test outcomes. For example, playback level (SPL) and background noise, respectively, can influence excitation pattern shapes and introduce undesired masking effects. Moreover, the presentation method can strongly influence perceived quality, because loudspeakers introduce distortions on their own and in conjunction with a listening room. These effects can introduce site dependencies. In short, although they have proven effective, existing subjective test procedures for audio codecs are clearly suboptimal. Recent research into more reliable tools for subjective codec evaluations has shown promise and is continuing. For example, Moulton Laboratories investigated [336], [337] the effectiveness of multifacet Rasch models [338] for improved reliability of subjective listening tests on high-quality audio codecs. The Rasch model [339] is a statistical analysis technique designed to remove the effects of local disturbances on test outcomes. The impact of Rasch analysis on the reliability of subjective audio codec evaluations is still under investigation. Meanwhile, the unreliability of subjective tests has motivated considerable research into development of automatic perceptual measurement schemes (e.g., [340]–[346], [186], [347]–[351]) that has ultimately led to the adoption of an international standard for perceptual quality measurement, ITU-R BS.1387 [352]. Experts do not consider the standardized algorithm to be a human subject replacement, however, and research into improved perceptual measurement schemes will continue (e.g., ITU-R JWP10-11Q). Automatic perceptual measurement of compressed high-fidelity audio quality is a fascinating topic that is treated in more detail elsewhere (e.g., [36] and [353]).

C. Subjective Evaluations of Two-Channel Standardized Codecs

The influence of site and subject dependencies on subjective listening tests can potentially invalidate direct comparisons between independent test results for different algorithms. Ideally, fair intercodec comparisons require that scores are obtained from a single site with the same test subjects. Soulodre, *et al.* conducted a formal ITU-R BS.1116-compliant [275] listening test that compared several standardized two-channel stereo codecs [354], including the MPEG-1 Layer 2 [17], the MPEG-1 Layer 3 [17], the MPEG-2 AAC [112], the Lucent Technologies PAC [16], and the Dolby AC-3 [268] codecs. In all, 17 algorithm/bit rate combinations were examined, using listening material deemed critical by experts.

The test results, reproduced in Table 2, clearly show eight performance classes. The AAC and AC-3 codecs at 128 and 192 kb/s, respectively, exhibited the best performance with mean difference grades better than −1.0. The MPEG-2 AAC algorithm at 128 kb/s, however, was the only codec that satisfied the quality requirements defined by ITU-R Rec. BS.1115 [355] for perceptual audio coding systems in broadcast applications, namely, that there not be any audio materials rated below −1.00. Overall, the ranking of the families from best to worst with respect to quality was AAC, PAC, MPEG-1 Layer 3, AC-3, MPEG-1 Layer 2, and ITIS (MPEG-1, LII, hardware implementation). The class three results can be interpreted to mean that bit rate increases of 32, 64, and 96 kb/s per stereo pair are required for the PAC, AC-3, and Layer 2 codec families, respectively, to match the output quality of the MPEG-2 AAC at 96 kb/s per stereo pair.

Table 2
Comparison of Standardized Two-Channel Algorithms (After [354])

Group	Algorithm	Rate (kbps)	Mean Diff. Grade	Transparent Items	Items Below −1.00
1	AAC	128	-0.47	1	0
	AC-3	192	-0.52	1	1
2	PAC	160	-0.82	1	3
3	PAC	128	-1.03	1	4
	AC-3	160	-1.04	0	4
	AAC	96	-1.15	0	5
	MP-1 L2	192	-1.18	0	5
4	IT IS	192	-1.38	0	6
5	MP-1 L3	128	-1.73	0	6
	MP-1 L2	160	-1.75	0	7
	PAC	96	-1.83	0	6
	IT IS	160	-1.84	0	6
6	AC-3	128	-2.11	0	8
	MP-1 L2	128	-2.14	0	8
	IT IS	128	-2.21	0	7
7	PAC	64	-3.09	0	8
8	IT IS	96	-3.32	0	8

Table 3
Comparison of Standardized 5.1-Channel Algorithms

Group	Algorithm	Rate (kbps)	Mean Diff. Grade
1	MP-2 BC	640	-0.51
2	AC-3	448	-0.93
	MP-2 BC	512	-0.99
3	AC-3	384	-1.17
	MP-2 BC	384	-1.73

D. Subjective Evaluations of 5.1-Channel Standardized Codecs

Multichannel perceptual audio coders are increasingly in demand for multimedia, cinema, and home theater applications. As a result, the European Broadcasting Union recently sponsored Deutsche Telekom Berkom in a formal subjective evaluation [356] that compared the output quality for real-time implementations of the 5.1 channel Dolby AC-3 and the matrixed 5.1-channel MPEG-2/BC Layer 2 algorithms at bit rates between 384 and 640 kb/s (Table 3). The tests adhered to the methodologies outlined in ITU BS.1116, and the five-channel listening environment was configured according to ITU-R Rec. BS.775 [357]. The resulting difference grades given in Table 3 represent averages of the mean grades reported for a collection of eight critical test items. None of the tested codec configurations satisfied "transparency." More sophisticated multichannel algorithms such as Lucent PAC and MPEG-2 AAC were not examined in this test because they were not considered to be sufficiently well established on the market [356].

Table 4
Audio Coding Standards and Applications

Algorithm	Sample Rates (kHz)	Channels	Bit Rates (kbps)	Applications	References
APT-X100	44.1	1	176.4	Cinema	[19]
ATRAC	44.1	2	256/ch	MiniDisc	[365]
Lucent PAC	44.1	1 - 5.1	128/stereo	DBA: 128/160 kbps	[306]
Dolby AC-2	44.1	2	256/ch	DBA	[313]
Dolby AC-3	44.1	1 - 5.1	32 – 384	Cinema, HDTV	[315]
MPEG-1, LI-III	32, 44.1, 48	1, 2	32 – 448	"MP3": LIII DBA: LII@256 kbps DBS: LII@224 kbps DCC: LI@384 kbps	[17]
MPEG-2/BC-LSF	32, 44.1, 48, 16, 22, 24	1 - 5.1	32 – 640	Cinema	[18]
MPEG-2/AAC		1 – 96	8 – 64 /ch	Internet/www, e.g., LiquidAudio™, atob™ audio	[112]
MPEG-4		1 -	200 bps – 64 kbps/ch	General	[222]

X. Conclusion

A. Summary of Applications for Commercial and International Standards

Current applications (Table 4) for embedded audio coding include DBA [358], [359], DBS [360], DVD [361], high-definition television (HDTV) [362], cinematic theater [363], and audio-on-demand over wide area networks such as the Internet [364]. Audio coding has also enabled miniaturization of digital audio storage media such as Compact MiniDisk [365] and DCC [366], [367]. With the advent of the ".MP3" audio format, which denotes audio files that have been compressed using the MPEG-1, Layer III algorithm, perceptual audio coding has become of central importance to over-network exchange of multimedia information, and has recently been integrated into several popular portable consumer audio playback devices that are specifically designed for web compatibility. In addition, DolbyNET, a version of the AC-3 algorithm, has been successfully integrated into streaming audio processors for delivery of audio on demand to the desktop Web browser.

B. Summary of Recent Research and Future Research Directions

The level of sophistication and high performance achieved by the standards listed in Table 4 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 [368]. Using primarily transform [369], subband (filter bank/wavelet) [370]–[374], and other [375]–[377] 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. Sinha and Johnston, for example, reported transparent CD quality at 64/32 kb/s for stereo/mono [373] sources. Other new algorithms include extended capacity for multichannel/multilanguage systems [363], [378], [379]. In addition to pursuing the usual goals of transparent compression at lower bit rates (below 64 kb/s/channel) with reduced complexity, minimal delay [380], and enhanced bit error robustness [401], an emerging trend for future research in audio coding is concerned with the development of algorithms that offer scalability [381]–[387]. Scalable algorithms will ultimately be used to accommodate the unique challenges associated with audio transmission over time-varying channels such as the packet-switched networks that compose the Internet, as well as time-varying wireless channels. Network-specific design considerations are also motivating research into joint source-channel coding [388] for audio over the Internet. Another emerging trend is one of convergence between low-rate audio coding algorithms and speech coders, which are increasingly embedding mechanisms to exploit perceptual irrelevancies [389], [390], [399], [400]. Research is also ongoing into potential improvements for the various perceptual coder building blocks, such as novel filter banks for low-delay coding and reduced pre-echo [391], [404] and new psychoacoustic signal analysis techniques [392], [393]. Researchers are also investigating new algorithms for tasks of peripheral interest to perceptual audio coding such as transform-domain signal modifications [394] and digital watermarking [395], [396]. Finally, considerable investigation is continuing into perceptual quality measurements for coder evaluations in terms of both subjective [336], [337] and objective methodologies. In fact, after a competition between and then ultimately a collaboration by several research teams, the ITU-R recently adopted an automatic perceptual measurement system, ITU-R BS-1387 [397], [402], [403] intended to assist in the tasks of codec selection, evaluation, and maintenance. Future research will continue in all of these areas.

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive comments and corrections.

References

[1] *Compact Disc Digital Audio System*, (IEC/ANSI) CEI-IEC-908, 1987.

[2] C. Todd, "A digital audio system for broadcast and prerecorded media," in *Proc. 75th Conv. Aud. Eng. Soc.*, Mar. 1984, preprint.
[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.*, Mar. 1986, preprint 2321.
[4] G. Theile, G. Stoll, and M. Link, "Low-bit rate coding of high quality audio signals," in *Proc. 82nd Conv. Aud. Eng. Soc.*, Mar. 1987, preprint 2432.
[5] K. Brandenburg, "OCF—A new coding algorithm for high quality sound signals," in *Proc. ICASSP-87*, May 1987, pp. 5.1.1–5.1.4.
[6] J. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 314–323, Feb. 1988.
[7] W.-Y. Chan and A. Gersho, "High fidelity audio transform coding with vector quantization," in *Proc. ICASSP-90*, May 1990, pp. 1109–1112.
[8] K. Brandenburg and J. D. Johnston, "Second generation perceptual audio coding: The hybrid coder," in *Proc. 88th Conv. Aud. Eng. Soc.*, Mar. 1990, preprint 2937.
[9] K. Brandenburg, J. Herre, J. D. Johnston, Y. Mahieux, and E. Schroeder, "ASPEC: Adaptive spectral entropy coding of high quality music signals," in *Proc. 90th Conv. Aud. Eng. Soc.*, Feb. 1991, preprint 3011.
[10] Y. F. Dehery, M. Lever, and P. Urcun, "A MUSICAM source codec for digital audio broadcasting and storage," in *Proc. ICASSP-91*, May 1991, pp. 3605–3608.
[11] M. Iwadare, A. Sugiyama, F. Hazu, A. Hirano, and T. Nishitiani, "A 128 kb/s hi-fi audio CODEC based on adaptive transform coding with adaptive block size MDCT," *IEEE J. Select. Areas Commun.*, pp. 138–144, Jan. 1992.
[12] 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.
[13] G. Stoll, S. Nielsen, and L. van de Kerkhof, "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.*, Oct. 1993, preprint 3697.
[14] J. B. Rault, P. Philippe, and M. Lever, "MUSICAM (ISO/MPEG audio) very low bit-rate coding at reduced sampling frequency," in *Proc. 95th Conv. Aud. Eng. Soc.*, Oct. 1993, preprint 3741.
[15] G. Stoll, G. Theile, S. Nielsen, A. Silzle, M. Link, R. Sedlmeyer, and A. Brefort, "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.*, Mar. 1993, preprint 3550.
[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, "Information technology—Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s—Part 3: Audio,", IS11172-3 1992 ("MPEG-1").
[18] ISO/IEC, JTC1/SC29/WG11 MPEG, "Information technology—Generic coding of moving pictures and associated audio—Part 3: Audio,", IS13818-3 1994 ("MPEG-2").
[19] F. Wylie, "Predictive or perceptual coding...apt-X and apt-Q," in *Proc. 100th Conv. Aud. Eng. Soc.*, May 1996, preprint 4200.
[20] P. Craven and M. Gerzon, "Lossless coding for audio discs," *J. Audio Eng. Soc.*, pp. 706–720, Sept. 1996.
[21] J. R. Stuart. (1995, June) A proposal for the high-quality audio application of high-density CD carriers. Technical Subcommittee Acoustic Renaissance for Audio. [Online] Available WWW: http://www.meridian.co.uk/ara/araconta.html.
[22] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[23] I. Witten, "Arithmetic coding for data compression," *Commun. ACM*, vol. 30, no. 6, pp. 520–540, June 1987.
[24] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theory*, vol. IT–23, pp. 337–343, May 1977.
[25] T. Welch, "A technique for high performance data compression," *IEEE Trans. Comput.*, vol. C-17, pp. 8–19, June 1984.
[26] N. Jayant, J. D. Johnston, and V. Shoham, "Coding of wideband speech," *Speech Commun.*, pp. 127–138, June 1992.
[27] 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.
[28] 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.
[29] N. Jayant, J. D. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, pp. 1385–1422, Oct. 1993.
[30] P. Noll, "Wideband speech and audio coding," *IEEE Commun. Mag.*, pp. 34–44, Nov. 1993.
[31] ——, "Digital audio coding for visual communications," *Proc. IEEE*, vol. 83, pp. 925–943, June 1995.
[32] K. Brandenburg, "Introduction to perceptual coding," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds., 1996, pp. 23–30.
[33] J. Johnston, "Audio coding with filter banks," in *Subband and Wavelet Transforms*, A. Akansu and M. J. T. Smith, Eds: Kluwer Academic, 1996, pp. 287–307.
[34] N. Gilchrist and C. Grewin, Eds., *Collected Papers on Digital Audio Bit-Rate Reduction*: Aud. Eng. Soc., 1996.
[35] *The Digital Signal Processing Handbook*, V. Madisetti and D. Williams, Eds., CRC Press, Boca Raton, FL, 1998, pp. 38.1–44.8.
[36] M. Kahrs and K. Brandenburg, Eds., *Applications of Digital Signal Processing to Audio and Acoustics*. Boston, MA: Kluwer Academic, 1998.
[37] H. Fletcher, "Auditory patterns," *Rev. Mod. Phys.*, pp. 47–65, Jan. 1940.
[38] D. D. Greenwood, "Critical bandwidth and the frequency coordinates of the Basilar membrane," *J. Acoust. Soc. Amer.*, pp. 1344–1356, Oct. 1961.
[39] J. Zwislocki, "Analysis of some auditory characteristics," in *Handbook of Mathematical Psychology*, R. Luce, R. Bush, and E. Galanter, Eds. New York: Wiley, 1965.
[40] B. Scharf, "Critical bands," in *Foundations of Modern Auditory Theory*. New York: Academic, 1970.
[41] R. Hellman, "Asymmetry of masking between noise and tone," *Percep. Psychphys.*, vol. 11, pp. 241–246, 1972.
[42] E. Zwicker and H. Fastl, *Psychoacoustics Facts and Models*. Berlin, Germany: Springer-Verlag, 1990.
[43] 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.
[44] M. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, pp. 1647–1652, Dec. 1979.
[45] J. Johnston, "Estimation of perceptual entropy using noise masking criteria," in *Proc. ICASSP-88*, May 1988, pp. 2524–2527.
[46] E. Terhardt, "Calculating virtual pitch," *Hearing Res.*, vol. 1, pp. 155–182, 1979.
[47] G. von Bekesy, *Experiments in Hearing*. New York: McGraw-Hill, 1960.
[48] D. Greenwood, "A cochlear frequency-position function for several species: 29 years later," *J. Acoust. Soc. Amer.*, vol. 87, pp. 2592–2605, June 1990.
[49] Boys Town National Research Hospital, Communication Engineering Laboratory. [Online] Available WWW: http://www.btnrh.boystown.org/cel/waves.htm; Department of Physiology at the University of Wisconsin–Madison. [Online]. Available WWW: http://www.neurophys.wisc.edu/animations/; Scuola Internazionale Superiore di Studi Avanzati/International School for Advanced Studies (SISSA/ISAS). [Online]. Available WWW: http:// www.sissa.it/bp/Cochlea/twlo.htm; and Ear Lab at Boston University. [Online]. Available: http://earlab.bu.edu/physiology/ mechanics.html.
[50] B. C. J. Moore, "Masking in the human auditory system," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds., 1996, pp. 9–19.
[51] B. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.*, vol. 47, pp. 103–138, 1990.
[52] B. C. J. Moore and B. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Amer.*, vol. 74, pp. 750–753, 1983.
[53] G. Gässler, "Uber die hörschwelle für schallereignisse mit verschieden breitem frequenzspektrum," *Acustica*, vol. 4, pp. 408–414, 1954.
[54] J. L. Hall, "Auditory psychophysics for coding applications," in *The Digital Signal Processing Handbook*, V. Madisetti and D. Williams, Eds. Boca Raton, FL: CRC Press, 1998, pp. 39.1–39.25.

[55] H. Fletcher and W. Munson, "Relation between loudness and masking," *J. Acoust. Soc. Amer.*, vol. 9, pp. 1–10, 1937.
[56] J. Egan and H. Hake, "On the masking pattern of a simple auditory stimulus," *J. Acoust. Soc. Amer.*, vol. 22, pp. 622–630, 1950.
[57] G. Miller, "Sensitivity to changes in the intensity of white noise and its relation to masking and loudness," *J. Acoust. Soc. Amer.*, vol. 19, pp. 609–619, 1947.
[58] J. L. Hall, "Asymmetry of masking revisited: Generalization of masker and probe bandwidth," *J. Acoust. Soc. Amer.*, vol. 101, pp. 1023–1033, Feb. 1997.
[59] N. Jayant, J. D. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, pp. 1385–1422, Oct. 1993.
[60] W. Jesteadt, S. Bacon, and J. Lehman, "Forward masking as a function of frequency, masker level, and signal delay," *J. Acoust. Soc. Amer.*, vol. 71, pp. 950–962, 1982.
[61] B. C. J. Moore, "Psychophysical tuning curves measured in simultaneous and forward masking," *J. Acoust. Soc. Amer.*, vol. 63, pp. 524–532, 1978.
[62] K. Brandenburg, "Perceptual coding of high quality digital audio," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds. Boston, MA: Kluwer Academic, 1998.
[63] P. Papamichalis, "MPEG audio compression: Algorithms and implementation," in *Proc. DSP 95 Int. Conf. DSP*, June 1995, pp. 72–77.
[64] N. Jayant and P. Noll, *Digital Coding of Waveforms Principles and Applications to Speech and Video*. Englewood Cliffs, NJ: Prentice-Hall, 1984.
[65] P. P. Vaidyanathan, "Quadrature mirror filter banks, M-band extensions, and perfect-reconstruction techniques," *IEEE Acoust., Speech, Signal Processing Mag.*, pp. 4–20, July 1987.
[66] ——, "Multirate digital filters, filter banks, polyphase networks, and applications: A tutorial," *Proc. IEEE*, vol. 78, pp. 56–93, Jan. 1990.
[67] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1983.
[68] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[69] A. Akansu and M. J. T. Smith, Eds., *Subband and Wavelet Transforms, Design and Applications*. Norwell, MA: Kluwer Academic, 1996.
[70] H. S. Malvar, *Signal Processing with Lapped Transforms*. Norwood, MA: Artech House, 1991.
[71] M. Vetterli and C. Herley, "Wavelets and filter banks," *IEEE Trans. Signal Processing*, vol. 40, pp. 2207–2232, Sept. 1992.
[72] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Mag.*, pp. 14–38, Oct. 1991.
[73] A. Akansu and R. Haddad, *Multiresolution Signal Decomposition: Transforms, Subbands, Wavelets*. San Diego, CA: Academic, 1992.
[74] G. Strang and T. Nguyen, *Wavelets and Filter banks*. Wellesley, MA: Wellesley-Cambridge, 1996.
[75] J. Johnston, S. Quackenbush, G. Davidson, K. Brandenburg, and J. Herre, "MPEG audio coding," in *Wavelet, Subband, and Block Transforms in Communications and Multimedia*, A. Akansu and M. Medley, Eds. Boston, MA: Kluwer Academic, 1999, ch. 7.
[76] K. Brandenburg, E. Eberlein, J. Herre, and B. Edler, "Comparison of filter banks for high quality audio coding," in *Proc. IEEE ISCAS*, 1992, pp. 1336–1339.
[77] H. J. Nussbaumer, "Pseudo QMF filter bank," *IBM Tech. Disclosure Bull.*, vol. 24, pp. 3081–3087, Nov. 1981.
[78] J. H. Rothweiler, "Polyphase quadrature filters: A new subband coding technique," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-83)*, May 1983, pp. 1280–1283.
[79] P. L. Chu, "Quadrature mirror filter design for an arbitrary number of equal bandwidth channels," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–33, pp. 203–218, Feb. 1985.
[80] J. Masson and Z. Picel, "Flexible design of computationally efficient nearly perfect QMF filter banks," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-85)*, Mar. 1985, pp. 14.7.1–14.7.4.
[81] R. Cox, "The design of uniformly and nonuniformly spaced pseudo QMF," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–34, pp. 1090–1096, Oct. 1986.
[82] D. Pan, "Digital audio compression," *Digital Tech. J.*, vol. 5, no. 2, pp. 28–40, 1993.
[83] H. Malvar, "Modulated QMF filter banks with perfect reconstruction," *Electron. Lett.*, vol. 26, pp. 906–907, June 1990.
[84] T. Ramstad, "Cosine modulated analysis–synthesis filter bank with critical sampling and perfect reconstruction," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-91)*, May 1991, pp. 1789–1792.
[85] R. Koilpillai and P. P. Vaidyanathan, "New results on cosine-modulated FIR filter banks satisfying perfect reconstruction," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-91)*, May 1991, pp. 1793–1796.
[86] ——, "Cosine-modulated FIR filter banks satisfying perfect reconstruction," *IEEE Trans. Signal Processing*, vol. SP–40, pp. 770–783, Apr. 1992.
[87] J. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1153–1161, Oct. 1986.
[88] H. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 969–978, June 1990.
[89] S. Cheung and J. Lim, "Incorporation of biorthogonality into lapped transforms for audio compression," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-95)*, May 1995, pp. 3079–3082.
[90] J. Princen, J. Johnson, and A. Bradley, "Subband/transform coding using filter bank designs based on time domain aliasing cancellation," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-87)*, May 1987, pp. 50.1.1–50.1.4.
[91] G. Smart and A. Bradley, "Filter bank design based on time-domain aliasing cancellation with nonidentical windows," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-94)*, May 1995, pp. III-185–III-188.
[92] B. Jawerth and W. Sweldens, "Biorthogonal smooth local trigonometric bases," *J. Fourier Anal. Appl.*, vol. 2, no. 2, pp. 109–133, 1995.
[93] G. Matviyenko, "Optimized local trigonometric bases," *Appl. Comput. Harmonic Anal.*, vol. 3, no. 4, pp. 301–323, 1996.
[94] A. Ferreira, "Convolutional effects in transform coding with TDAC: An optimal window," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 104–114, Mar. 1996.
[95] H. Malvar, "Biorthogonal and nonuniform lapped transforms for transform coding with reduced blocking and ringing artifacts," *IEEE Trans. Signal Processing*, vol. 46, pp. 1043–1053, Apr. 1998.
[96] C. Herley, "Boundary filters for finite-length signals and time-varying filter banks," *IEEE Trans. Circuits Syst. II*, vol. 42, pp. 102–114, Feb. 1995.
[97] C. Herley, J. Kovacevic, and K. Ramchandran, "Tilings of the time-frequency plane: Construction of arbitrary orthogonal bases and fast tiling algorithms," *IEEE Trans. Signal Processing*, vol. 41, pp. 3341–3359, 1993.
[98] I. Sodagar, K. Nayebi, and T. Barnwell, "Time-varying filter banks and wavelets," *IEEE Trans. Signal Processing*, vol. 42, pp. 2983–2996, Nov. 1994.
[99] R. de Queiroz, "Time-varying lapped transforms and wavelet packets," *IEEE Trans. Signal Processing*, vol. 41, pp. 3293–3305, 1993.
[100] P. Duhamel, Y. Mahieux, and J. Petit, "A fast algorithm for the implementation of filter banks based on time domain aliasing cancellation," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-91)*, May 1991, pp. 2209–2212.
[101] D. Sevic and M. Popovic, "A new efficient implementation of the oddly-stacked princen-bradley filter bank," *IEEE Signal Processing Lett.*, vol. 1, pp. 166–168, Nov. 1994.
[102] C.-M. Liu and W.-C. Lee, "A unified fast algorithm for cosine modulated filter banks in current audio coding standards," in *Proc. 104th Conv. Aud. Eng. Soc.*, 1998, preprint 4729.
[103] H.-C. Chiang and J.-C. Liu, "Regressive implementations for the forward and inverse MDCT in MPEG audio coding," *IEEE Signal Processing Lett.*, vol. 3, pp. 116–118, Apr. 1996.
[104] C. Jakob and A. Bradley, "Minimizing the effects of subband quantization of the time domain aliasing cancellation filter bank," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-96)*, May 1996, pp. 1033–1036.
[105] B. Edler, "Codierung von audiosignalen mit überlappender transformation und adaptiven fensterfunktionen," *Frequenz*, pp. 252–256, 1989.
[106] S. Shlien, "The modulated lapped transform, its time-varying forms, and its applications to audio coding standards," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 359–366, July 1997.

[107] T. Vaupel, "Ein beitrag zur transformationscodierung von audiosignalen unter verwendung der methode der 'time domain aliasing cancellation (TDAC)' und einer signalkompandierung in zeitbereich," Ph.D. dissertation, Univ. Duisburg, Duisburg, Germany, Apr. 1991.
[108] M. Link, "An attack processing of audio signals for optimizing the temporal characteristics of a low bit-rate audio coding system," in *Proc. 95th Conv. Aud. Eng. Soc.*, 1993, preprint 3696.
[109] K. Akagiri, "Technical description of Sony preprocessing," SO/IEC JTC1/SC29/WG11 MPEGI, Input Doc., 1994.
[110] J. Herre and J. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)," in *Proc. 101st Conv. Aud. Eng. Soc.*, 1996, preprint 4384.
[111] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, and M. Dietz, "MPEG-2 advanced audio coding," in *Proc. 101st Conv. Aud. Eng. Soc.*, 1996, preprint.
[112] ISO/IEC, JTC1/SC29/WG11 MPEG, "Generic coding of moving pictures and associated audio—Audio (non backward compatible coding, NBC)," JTC1/SC29/WG11 MPEG, Committee Draft 13 818-7 1996 ("MPEG-2 NBC/AAC").
[113] D. Krahe, "New source coding method for high quality digital audio signals," *NTG Fachtagung Hoerrundfunk*, pp. S.371–S.381, 1985.
[114] ——, "Grundlagen eines verfahrens zur dataenreduktion bei qualitativ hochwertigen, digitalen audiosignalen auf basis einer adaptiven transformationscodierung unter berucksightigung psychoakustischer phanomene," Ph.D. dissertation, Unive. Duisburg, Duisburg, Germany, 1988.
[115] K. Brandenburg, "High quality sound coding at 2.5 bits/sample," in *Proc. 84th Conv. Aud. Eng. Soc.*, Mar. 1988, preprint 2582.
[116] ——, "OCF: Coding high quality audio with data rates of 64 kbit/sec," in *Proc. 85th Conv. Aud. Eng. Soc.*, Mar. 1988, preprint 2723.
[117] J. Johnston, "Perceptual transform coding of wideband stereo signals," in *Proc. ICASSP-89*, May 1989, pp. 1993–1996.
[118] Y. Mahieux, Y. Mahieux, J. Petit, and A. Charbonnier, "Transform coding of audio signals using correlation between successive transform blocks," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-89)*, May 1989, pp. 2021–2024.
[119] Y. Mahieux and J. Petit, "Transform coding of audio signals at 64 kbits/sec," in *Proc. Globecom'90*, Nov. 1990, pp. 405.2.1–405.2.5.
[120] A. Sugiyama, F. Hazu, M. Iwadare, and T. Nishitani, "Adaptive transform coding with an adaptive block size (ATC-ABS)," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-90)*, May 1990, pp. 1093–1096.
[121] M. Paraskevas and J. Mourjopoulos, "A differential perceptual audio coding method with reduced bitrate requirements," *IEEE Trans. Speech Audio Processing*, pp. 490–503, Nov. 1995.
[122] D. Schulz, "Improving audio codecs by noise substitution," *J. Audio Eng. Soc.*, pp. 593–598, July/Aug 1996.
[123] W. Chan and A. Gersho, "Constrained-storage vector quantization in high fidelity audio transform coding," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-91)*, May 1991, pp. 3597–3600.
[124] ——, "Constrained-storage quantization of multiple vector sources by codebook sharing," *IEEE Trans. Commun.*, pp. 11–13, Jan. 1991.
[125] N. Iwakami, T. Moriya, and S. Miki, "High-quality audio-coding at less than 64 kbit/s by using transform-domain weighted interleave vector quantization (TWINVQ)," in *Proc. ICASSP-95*, May 1995, pp. 3095–3098.
[126] N. Iwakami and T. Moriya, "Transform domain weighted interleave vector quantization (TwinVQ)," in *Proc. 101st Conv. Aud. Eng. Soc.*, Nov. 1996, preprint 4377.
[127] ISO/IEC, JTC1/SC29/WG11 (MPEG) document N2011, "Results of AAC and TwinVQ tool comparative tests,", San Jose, CA, 1998.
[128] T. Moriya, N. Iwakami, K. Ikeda, and S. Miki, "Extension and complexity reduction of TWINVQ audio coder," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-96)*, May 1996, pp. 1029–1032.
[129] K. Ikeda, T. Moriya, and N. Iwakami, "Error protected TwinVQ audio coding at less than 64 kbit/s," in *Proc. IEEE Speech Coding Workshop*, 1995, pp. 33–34.
[130] A. Charbonnier and J. P. Petit, "Sub-band ADPCM coding for high quality audio signals," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-88)*, May 1988, pp. 2540–2543.
[131] P. Voros, "High-quality sound coding within 2x64 kbit/s using instantaneous dynamic bit-allocation," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-88)*, May 1988, pp. 2536–2539.
[132] D.-H. Teh, A.-P. Tan, and S.-N. Koh, "Subband coding of high-fidelity quality audio signals at 128 kb/s," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-92)*, May 1990, pp. II-197–II-200.
[133] G. Stoll, M. Link, and G. Theile, "Masking-pattern adapted subband coding: Use of the dynamic bit-rate margin," in *Proc. 84th Conv. Aud. Eng. Soc.*, Mar. 1988, preprint 2585.
[134] R. N. J. Veldhuis, "Subband coding of digital audio signals without loss of quality," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-89)*, May 1989, pp. 2009–2012.
[135] D. Wiese and G. Stoll, "Bitrate reduction of high quality audio signals by modeling the ear's masking thresholds," in *Proc. 89th Conv. Aud. Eng. Soc.*, Sept. 1990, preprint 2970.
[136] "ISO MPEG/audio test report," Swedish Broadcasting Corp., Stockholm, Sweden, July 1990.
[137] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1992.
[138] S. Boland and M. Deriche, "New results in low bitrate audio coding using a combined harmonic-wavelet representation," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-97)*, Apr. 1997, pp. 351–354.
[139] A. Pena, C. Serantes, and N. Prelcic, "ARCO (Adaptive Resolution CODec): A hybrid approach to perceptual audio coding," in *Proc. 100th Conv. Aud. Eng. Soc.*, May 1996, preprint 4178.
[140] N. Gonzalez-Prelcic, S. Gonzalez, and A. Pena, "Considerations on the performance of filter design methods for wavelet packet audio decomposition," in *Proc. 100th Conv. Aud. Eng. Soc.*, May 1996, preprint 4235.
[141] N. Prelcic and A. Pena, "An adaptive tree search algorithm with application to multiresolution based perceptive audio coding," in *Proc. IEEE Int. Symp. Time-Frequency and Time-Scale Analysis*, 1996, pp. 117–120.
[142] A. Pena, C. Serantes, and N. Gonzalez-Prelcic, "New improvements in ARCO (Adaptive Resolution CODec)," in *Proc. 102nd Conv. Aud. Eng. Soc.*, Mar. 1997, preprint 4419.
[143] M. Black and M. Zeytinoglu, "Computationally efficient wavelet packet coding of wideband stereo audio signals," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-95)*, May 1995, pp. 3075–3078.
[144] P. Kudumakis and M. Sandler, "On the performance of wavelets for low bit rate coding of audio signals," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-95)*, May 1995, pp. 3087–3090.
[145] ——, "Wavelets, regularity, complexity, and MPEG-audio," in *Proc. 99th Conv. Aud. Eng. Soc.*, Oct. 1995, preprint 4048.
[146] ——, "On the compression obtainable with four-tap wavelets," *IEEE Signal Processing Lett.*, pp. 231–233, Aug. 1996.
[147] S. Boland and M. Deriche, "High quality audio coding using multipulse LPC and wavelet decomposition," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-95)*, May 1995, pp. 3067–3069.
[148] ——, "Audio coding using the wavelet packet transform and a combined scalar-vector quantization," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-96)*, May 1996, pp. 1041–1044.
[149] W. Dobson, J. Yang, K. Smart, and F. Guo, "High quality low complexity scalable wavelet audio coding," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-97)*, Apr. 1997, pp. 327–330.
[150] Z. Lu and W. Pearlman, "An efficient, low-complexity audio coder delivering multiple levels of quality for interactive applications," in *Proc. IEEE Signal Processing Soc. Workshop Multimedia Signal Processing*, Dec. 1998.
[151] R. Coifman and M. Wickerhauser, "Entropy based algorithms for best basis selection," *IEEE Trans. Inform. Theory*, vol. 38, pp. 712–718, Mar. 1992.
[152] M. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*. Wellesley, MA: A. K. Peters, 1994.
[153] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. 27, pp. 379–423–623–656, 1948.
[154] R. Hedges, "Hybrid wavelet packet analysis," in *Proc. 31st Asilomar Conf. on Sig., Sys., and Comp.*, Oct. 1997, pp. 1254–1258.
[155] R. Hedges and D. Cochran, "Hybrid wavelet packet analysis," in *Proc. IEEE SP Int. Symp. Time-Frequency and Time-Scale Analysis*, Oct. 1998, pp. 221–224.

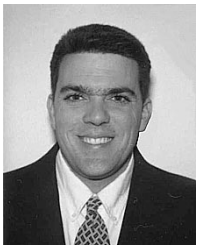
[156] ——, "Hybrid wavelet packet analysis: A top down approach," in *Proc. 32nd Asilomar Conf. Sig., Sys., and Comp.*, vol. 2, Nov. 1998, pp. 1381–1385.
[157] D. Sinha and A. Tewfik, "Low bit rate transparent audio compression using a dynamic dictionary and optimized wavelets," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-93)*, May 1993, pp. I-197–I-200.
[158] ——, "Low bit rate transparent audio compression using an adapted wavelets," *IEEE Trans. Signal Processing*, vol. 41, pp. 3463–3479, Dec. 1993.
[159] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Commun. Pure Appl. Math.*, pp. 909–996, Nov. 1988.
[160] A. Tewfik and M. Ali, "Enhanced wavelet based audio coder," in *Conf. Rec. 27th Asilomar Conf. Sig. Sys., and Comp.*, Nov. 1993, pp. 896–900.
[161] P. Srinivasan and L. Jamieson, "High quality audio compression using an adaptive wavelet packet decomposition and psychoacoustic modeling," *IEEE Trans. Signal Processing*, vol. 46, pp. 1085–1093, Apr. 1998.
[162] P. Srinivasan, "Speech and wideband audio compression using filter banks and wavelets," Ph. D. dissertation, Purdue Univ., West Lafayette, IN, May 1997.
[163] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.
[164] M. Erne, G. Moschytz, and C. Faller, "Best wavelet-packet bases for audio coding using perceptual and rate-distortion criteria," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-99)*, Mar. 1999, pp. 909–912.
[165] M. Erne and G. Moschytz, "Perceptual and near-lossless audio coding based on a signal-adaptive wavelet filterbank," in *Proc. 106th Conv. Aud. Eng. Soc.*, May 1999, preprint 4934.
[166] P. Philippe, M. Lever, J.-B. Rault, F. Moreau de Saint Martin, and J. Soumagne, "A relevant criterion for the design of wavelet filters in high-quality audio coding," in *Proc. 98th Conv. Aud. Eng. Soc.*, Feb. 1995, preprint 3948.
[167] P. Philippe, F. Moreau de Saint-Martin, and L. Mainard, "On the choice of wavelet filters for audio compression," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-95)*, May 1995, pp. 1045–1048.
[168] O. Rioul and P. Duhamel, "A Remez exchange algorithm for orthonormal wavelets," *IEEE Trans. Circuits Syst. II*, pp. 550–560, Aug. 1994.
[169] P. Onno and C. Guillemot, "Tradeoffs in the design of wavelet filters for image compression," in *Proc. VCIP*, Nov. 1993, pp. 1536–1547.
[170] F. Moreau de Saint-Martin, A. Cohen, and P. Sioha, "A measure of near orthogonality of PR biorthogonal filter banks," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-95)*, May 1995, pp. 1480–1483.
[171] P. Philippe, F. Moreau de Saint-Martin, M. Lever, and J. Soumagne, "Optimal wavelet packets for low-delay audio coding," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-96)*, May 1996, pp. 550–553.
[172] S. Kirkpatrick, C. Gelatt, Jr., and M. Vecchi, "Optimization by simulated annealing," *Science*, pp. 671–680, May 1983.
[173] M. J. T. Smith and T. Barnwell, "Exact reconstruction techniques for tree-structered subband coders," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 434–441, June 1986.
[174] H. Caglar *et al.*, "Statistically optimized PR-QMF design," in *Proc. SPIE Vis. Commun. Image Proc.*, Nov. 1991, pp. 86–94.
[175] P. Philippe, F. Moreau de Saint-Martin, and M. Lever, "Wavelet packet filterbanks for low time delay audio coding," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 310–322, May 1999.
[176] P. P. Vaidyanathan and T. Chen, "Statistically optimal synthesis banks for subband coders," in *Proc. 28th Asilomar Conf. Sig., Sys., and Comp.*, Nov. 1994.
[177] B. Chen, C.-W. Lin, and Y.-L. Chen, "Optimal signal reconstruction in noisy filterbanks: Multirate kalman synthesis filtering approach," *IEEE Trans. Signal Processing*, vol. 43, pp. 2496–2504, Nov. 1995.
[178] J. Kovacevic, "Subband coding systems incorporating quantizer models," *IEEE Trans. Image Process.*, pp. 543–553, May 1995.
[179] R. Haddad and K. Park, "Modeling, analysis, and optimum design of quantized M-band filterbanks," *IEEE Trans. Signal Processing*, vol. 43, pp. 2540–2549, Nov. 1995.
[180] A. Delopoulos and S. Kollias, "Optimal filterbanks for signal reconstruction from noisy subband components," *IEEE Trans. Signal Processing*, vol. 44, pp. 212–224, Feb. 1996.
[181] K. Gosse, F. Moreau de Saint-Martin, and P. Duhamel, "Filterbank design for minimum distortion in the presence of subband quantization," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-96)*, May 1996, pp. 1491–1494.
[182] K. Gosse and P. Duhamel, "Perfect reconstruction versus MMSE filterbanks in source coding," *IEEE Trans. Signal Processing*, vol. 45, pp. 2188–2202, Sept. 1997.
[183] K. Gosse, O. Pothier, and P. Duhamel, "Optimizing the synthesis filter bank in audio coding for minimum distortion using a frequency weighted psychoacoustic criterion," in *Proc. IEEE ASSP Workshop App. Signal Processing to Audio and Acoustics*, 1995, pp. 191–194.
[184] K. Gosse, F. Moreau de Saint-Martin, X. Durot, P. Duhamel, and J.-B.. Rault, "Subband audio coding with synthesis filters minimizing a perceptual distortion," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-97)*, May 1997, pp. 347–50.
[185] X. Durot and J.-B. Rault, "A new noise injection model for audio compression algorithms," in *Proc. 101st Conv. Aud. Eng. Soc.*, Nov. 1996, preprint 4374.
[186] C. Colomes, M. Lever, and J. Rault, "A perceptual model applied to audio bit-rate reduction," *J. Aud. Eng. Soc.*, pp. 233–240, Apr. 1995.
[187] K. Hamdy, M. Ali, and A. Tewfik, "Low bit rate high quality audio coding with combined harmonic and wavelet representations," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-96)*, May 1996, pp. 1045–1048.
[188] D. Thomson, "Spectrum estimation and harmonic analysis," *Proc. IEEE*, vol. 70, pp. 1055–1096, Sept. 1982.
[189] R. McAulay and T. Quatieri, "Speech analysis synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 744–754, Aug. 1986.
[190] M. Ali, "Adaptive signal representation with application in audio coding," Ph.D. dissertation, Univ. of Minnesota, Mar. 1996.
[191] O. Alkin and H. Calgar, "Design of efficient M-band coders with linear phase and perfect-reconstruction properties," *IEEE Trans. Signal Processing*, vol. 43, pp. 1579–1589, July 1995.
[192] "SQAM-sound quality assessment material: Recordings for subjective tests," EBU, Tech. Doc. 3253 (includes SQAM Compact Disc), 1988.
[193] A. Pena, N. Gonzalez-Prelcic, and C. Serantes, "A flexible tiling of the time axis for adaptive wavelet packet decompositions," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-97)*, Apr. 1997, pp. 2137–2140.
[194] E. Terhardt *et al.*, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *J. Acoust. Soc. Amer.*, vol. 71, pp. 679–688, Mar. 1982.
[195] C. Serantes, A. Pena, and N. Gonzalez-Prelcic, "A fast noise-scaling algorithm for uniform quantization in audio coding schemes," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-97)*, Apr. 1997, pp. 339–342.
[196] A. Pena, "A suggested auditory information environment to ease the detection and minimization of subjective annoyance in perceptive-based systems," in *Proc. 98th Conv. Aud. Eng. Soc.*, Paris, France, 1995, preprint 4019.
[197] A. Casal, C. Serantes, N. Gonzalez-Prelcic, and A. Pena, "Testing a flexible time-frequency mapping for high frequencies in TARCO (Tonal Adaptive Resolution COdec)," in *Proc. 104th Conv. Aud. Eng. Soc.*, Amsterdam, The Netherlands, May 1998, preprint 4676.
[198] J. Princen, "The design of nonuniform modulated filterbanks," in *Proc. IEEE Int. Symp. on Time-Frequency and Time-Scale Analysis*, Oct. 1994, pp. 112–115.
[199] P. Monta and S. Cheung, "Low rate audio coder with hierarchical filterbanks," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-94)*, May 1994, pp. II-209–II-212.
[200] L. Mainard and M. Lever, "A bi-dimensional coding scheme applied to audio bit rate reduction," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-96)*, May 1994, pp. 1017–1020.
[201] J. Princen and J. Johnston, "Audio coding with signal adaptive filterbanks," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-95)*, May 1995, pp. 3071–3074.
[202] M. Purat and P. Noll, "A new orthonormal wavelet packet decomposition for audio coding using frequency-varying modulated lapped transforms," in *IEEE ASSP Workshop Applic. of Signal Processing to Aud. and Acoustics*, Oct. 1995, Session 8.
[203] C. Creusere and S. Mitra, "Efficient audio coding using perfect reconstruction noncausal IIR filter banks," *IEEE Trans. Speech Audio Processing*, pp. 115–123, Mar. 1996.

[204] P. Hedelin, "A tone-oriented voice-excited vocoder," in *Proc. IEEE Int. Conf. Acoustics Speech, and Signal Processing (ICASSP-81)*, Mar. 1981, pp. 205–208.
[205] L. Almeida, "Nonstationary spectral modeling of voiced speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 374–390, June 1983.
[206] S. Levine and J. Smith, "A sines+transients+noise audio representation for data compression and time/pitch scale modifications," in *Proc. 105th Conv. Aud. Eng. Soc.*, Sept. 1998, preprint 4781.
[207] T. Verma and T. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *Proc. IEEE Int. Conf. Acoustics Speech, and Signal Processing (ICASSP-99)*, Mar. 1999, pp. 981–984.
[208] S. Levine and J. O. Smith III, "A switched parametric and transform audio coder," in *Proc. IEEE Int. Conf. Acoustics Speech, and Signal Processing (ICASSP-99)*, Mar. 1999, pp. 985–988.
[209] S. Levine, "Audio representations for data compression and compressed domain processing," Ph.D. dissertation, Stanford Univ., Stanford, CA, Dec. 1998.
[210] B. Edler, "Technical description of the MPEG-4 audio coding proposal from University of Hannover and Deutsche Bundespost Telekom," ISO/IEC, JTC1/SC29/WG11 MPEG95/MO414, Oct. 1995.
[211] B. Edler and H. Purnhagen, "Technical description of the MPEG-4 audio coding proposal from University of Hannover and Deutschet Telekom AG," ISO/IEC, JTC1/SC29/WG11 MPEG96/MO632, Jan. 1996.
[212] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones," *J. Aud. Eng. Soc.*, pp. 497–516, June 1992.
[213] ——, "Speech analysis/synthesis and modification using and analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. Speech Audio Processing*, pp. 389–406, Sept. 1997.
[214] X. Serra and J. O. Smith III, "Spectral modeling and synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Comput. Mus. J.*, pp. 12–24, Winter 1990.
[215] F. Baumgarte, C. Ferekidis, and H. Fuchs, "A nonlinear psychoacoustic model applied to the ISO MPEG layer 3 coder," in *Proc. 99th Conv. Aud. Eng. Soc.*, New York, Oct. 1995, preprint 4087.
[216] B. Edler , H. Purnhagen, and C. Ferekidis, "ASAC—Analysis/synthesis audio codec for very low bit rates," in *Proc. 100th Conv. Aud. Eng. Soc.*, May 1996, preprint 4179.
[217] ISO/IEC, JTC1/SC29/WG11, "MPEG-4 audio test results (MOS tests)," ISO/IEC, Munich, Germany, JTC1/SC29/WG11/N1144, Jan. 1996.
[218] ISO/IEC, JTC1/SC29/WG11, "Report of the *Ad Hoc* Group on the evaluation of new audio submissions to MPEG-4," ISO/IEC, Munich, Germany, JTC1/SC29/WG11/MPEG96/M0680, Jan. 1996.
[219] H. Purnhagen, B. Edler, and C. Ferekidis, "Object-based analysis/synthesis audio coder for very low-bit rates," in *Proc. 104th Conv. Aud. Eng. Soc.*, May 1998, preprint 4747.
[220] H. Purnhagen, "Proposal of a core experiment for extended 'harmonic and individual lines plus noise' tools for the parametric audio coder core," ISO/IEC, JTC1/SC29/WG11 MPEG97/2480, July 1997.
[221] H. Purnhagen and B. Edler, "Check phase results of core experiment on extended 'harmonic and individual lines plus noise'," ISO/IEC, JTC1/ SC29/WG11 MPEG97/2795, Oct. 1997.
[222] ISO/IEC, JTC1/SC29/WG11, "MPEG-4 audio committee draft 14 496-3," ISO/IEC, Available WWW: http://www.tnt.uni-hannover.de/project/mpeg/audio/documents, Oct. 1997.
[223] B. Feiten , R. Schwalbe, and F. Feige, "Dynamically scalable audio internet transmission," in *Proc. 104th Conv. Aud. Eng. Soc.*, May 1998, preprint 4686.
[224] J. Chowing, "The synthesis of complex audio spectra by means of frequency modulation," *J. Aud. Eng. Soc.*, pp. 526–529, Sept. 1973.
[225] B. Winduratna, "FM analysis/synthesis based audio coding," in *Proc. 104th Conv. Aud. Eng. Soc.*, May 1998, preprint 4746.
[226] A. J. S. Ferreira, "Perceptual coding of harmonic signals," in *Proc. 100th Conv. Aud. Eng. Soc.*, May 1996, preprint 4746.
[227] B. Edler and L. Contin, "MPEG-4 audio test results (MOS test),", ISO/IEC JTC1/SC29/WG11 N1144, Jan. 1996.
[228] B. Edler and H. Purnhagen, "Concepts for hybrid audio coding schemes based on parametric techniques," in *Proc. 105th Conv. Aud. Eng. Soc.*, 1998, preprint 4808.
[229] J. Saunders, "Real time discrimination of broadcast speech/music," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-96)*, May 1996, pp. 993–996.
[230] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-98)*, May 1998.
[231] B. Edler, "Very low bit rate audio coding development," in *Proc. 14th Aud. Eng. Soc. Int. Conf.*, June 1997.
[232] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.
[233] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, June 1996.
[234] A. Spanias, "Speech coding: A tutorial review," *Proc. IEEE*, vol. 82, pp. 1541–1582, Oct. 1994.
[235] S. Singhal, "High quality audio coding using multipulse LPC," in *Proc. ICASSP-90*, May 1990, pp. 1101–1104.
[236] X. Lin , R. Salami, and R. Steele, "High quality audio coding using analysis-by-synthesis technique," in *Proc. ICASSP-91*, May 1991, pp. 3617–3620.
[237] S. Boland and M. Deriche, "Hybrid LPC And discrete wavelet transform audio coding with a novel bit allocation algorithm," in *Proc. ICASSP-98*, May 1998, pp. 3657–3660.
[238] W. Chang and C. Want, "A masking-threshold-adapted weighting filter for excitation search," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 124–132, Mar. 1996.
[239] W.-W Chang , D.-Y. Wang , and L.-W. Wang , "Audio coding using sinusoidal excitation representation," *Proc. ICASSP-97*, pp. 311–314, Apr. 1997.
[240] A. Oppenheim, D. Johnson, and K. Steiglitz, "Computation of spectra with unequal resolution using the fast Fourier transform," *Proc. IEEE*, vol. 59, pp. 299–301, Feb. 1971.
[241] A. Oppenheim and D. Johnson, "Discrete representation of signals," *Proc. IEEE*, vol. 60, pp. 681–691, June 1972.
[242] H. Strube, "Linear prediction on a warped frequency scale," *J. Acoust. Soc. Amer.*, vol. 68, no. 4, pp. 1071–1076, Oct. 1980.
[243] E. Kruger and H. Strube, "Linear prediction on a warped frequency scale," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1529–1531, Sept. 1988.
[244] J. O. Smith and J. Abel, "The bark bilinear transform," in *Proc. IEEE Workshop App. Signal Processing to Audio and Electroacoustics*, Oct. 1995, Available WWW: http://www-ccrma.stanford.edu/~jos/.
[245] A. Harma, U. Laine, and M. Karjalainen, "Warped linear prediction (WLP) in audio coding," in *Proc. NORSIG'96*, Sept. 1996, pp. 367–370.
[246] ——, "An experimental audio codec based on warped linear prediction of complex valued signals," in *Proc. ICASSP-97*, Apr. 1997, pp. 323–326.
[247] ——, "WLPAC—A perceptual audio codec in a nutshell," in *Proc. 102nd Audio Eng. Soc. Conv.*, Munich, Germany, 1997, preprint 4420.
[248] A. Harma , M. Vaalgamaa, and U. Laine, "A warped linear predictive stereo codec using temporal noise shaping," in *Proc. ICASSP-98*, May 1998.
[249] "Information technology—Coding of moving pictures and associated audio for digital storage media at up to About 1.5 Mbit/s-IS 11 172-3 (audio)," ISO/IEC, JTC1/SC29, 1992.
[250] G. Stoll, "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.*, Berlin, Germany, 1993, preprint 3550.
[251] B. Grill, J. Herre, K. Brandenburg, E. Eberlein, J. Koller, and J. Muller, "Improved MPEG-2 audio multi-channel encoding," in *Proc. 96th Audio Eng. Soc. Conv.*, Amsterdam, The Netherlands, 1994, preprint 3865.
[252] "Information technology—Generic coding of moving pictures and associated audio information-DIS 13 818-3 (audio)," ISO/IEC, JTC1/SC29, 1994.
[253] W. Th. ten Kate , "Compatibility matrixing of multi-channel bit rate reduced audio signals," in *Proc. 96th Audio Eng. Soc. Conv.*, Amsterdam, The Netherlands, 1994, preprint 3792.
[254] 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.
[255] S. Shlien, "Guide to MPEG-1 audio standard," *IEEE Trans. Broadcast.*, pp. 206–218, Dec. 1994.

[256] D. Pan, "A tutorial on MPEG/audio compression," *IEEE Mult. Med.*, pp. 60–74, Summer 1995.
[257] P. Noll, "MPEG digital audio coding," *IEEE Signal Processing Mag.*, pp. 59–81, Sept. 1997.
[258] R. Storey, "ATLANTIC: Advanced television at low bitrates networked transmission over integrated communication systems," *ACTS Common European Newsletter*, Feb. 1997.
[259] N. Gilchrist, "ATLANTIC audio: Preserving technical quality during low bit rate coding and decoding," in *Proc. 104th Audio Eng. Soc. Conv.*, Amsterdam, The Netherlands, May 1998, preprint 4694.
[260] P. Lauber and N. Gilchrist, "ATLANTIC audio: Switching layer 3 signals," in *Proc. 104th Audio Eng. Soc. Conv.*, Amsterdam, The Netherlands, May 1998, preprint 4738.
[261] S. Ritscher and U. Felderhoff, "Cascading of different audio codecs," in *Proc. 100th Audio Eng. Soc. Conv.*, Copenhagen, Denmark, May 1996, preprint 4174.
[262] J. Fletcher, "ISO/MPEG layer 2—Optimum re-encoding of decoded audio using A MOLE signal," in *Proc. 104th Audio Eng. Soc. Conv.*, Amsterdam, The Netherlands, 1998, see also http://www.bbc.co.uk/atlantic.
[263] W. R. T. ten Kate, "Maintaining audio quality in cascaded psychoacoustic coding," in *Proc. 101st Audio Eng. Soc. Conv.*, Los Angeles, CA, Nov. 1996, preprint 4387.
[264] "Basic audio quality requirements for digital audio bit rate reduction systems for broadcast emission and primary distribution,", ITU-R Document TG10-2/3-E only, Oct. 1991.
[265] ISO/IEC, 13 818-7, "Information technology—Generic coding of moving pictures and associated audio—Part 7: Advanced audio coding,", 1997.
[266] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, and M. Dietz, "ISO/IEC MPEG-2 advanced audio coding," in *Proc. 101st Audio Eng. Soc. Conv.*, Los Angeles, CA, 1996, preprint 4382.
[267] ——, "ISO/IEC MPEG-2 advanced audio coding," *J. Audio Eng. Soc.*, pp. 789–813, Oct. 1997.
[268] L. Fielder, M. Bosi, G. Davidson, M. Davis, C. Todd, S. Vernon, and L. Fielder , "AC-2 and AC-3: Low-complexity transform-based audio coding," *Collected Papers Digital Audio Bit-Rate Reduction*, pp. 54–72, 1996.
[269] G. Davidson , L. Fielder, and B. Link, "Parametric bit allocation in a perceptual audio coder," in *Proc. 97th Audio Eng. Soc. Conv.*, Nov. 1994, preprint 3921.
[270] H. Fuchs, "Improving joint stereo audio coding by adaptive interchannel prediction," in *Proc. 1993 IEEE ASSP Workshop Apps. of Signal Processing to Aud. and Acoustics*, 1993.
[271] ——, "Improving MPEG audio coding by backward adaptive linear stereo prediction," in *Proc. 99th Conv. Aud. Eng. Soc.*, Oct. 1995, preprint 4086.
[272] S. Quackenbush, "Noiseless coding of quantized spectral components in MPEG-2 advanced audio coding," in *IEEE ASSP Workshop Apps. of Signal Processing to Aud. and Acoustics*, Mohonk, 1997.
[273] J. Johnston, J. Herre, M. Davis, and U. Gbur, "MPEG-2 NBC audio–stereo and multichannel coding methods," in *Proc. 101st Audio Eng. Soc. Conv.*, Los Angeles, CA, 1996, preprint 4383.
[274] ISO/IEC, JTC1/SC29/WG11 N1420, "Overview of the report on the formal subjective listening tests of MPEG-2 AAC multichannel audio coding,", Nov. 1996.
[275] "Methods for subjective assessment of small impairments in audio systems including multichannel sound systems,", ITU-R BS 1116, 1994.
[276] D. Kirby and K. Watanabe, "Formal subjective testing of the MPEG-2 NBC multichannel coding algorithm," in *Proc. 102nd Audio Eng. Soc. Conv.*, Munich, Germany, 1997, preprint 4418.
[277] S. Quackenbush and Y. Toguri, "Revised report on complexity of MPEG-2 AAC tools," ISO/IEC, JTC1/SC29/WG11 N2005, Feb. 1998.
[278] L. Yin , M. Suonio, and M. Vaananen, "A new backward predictor for MPEG audio coding," in *Proc. 103rd Audio Eng. Soc. Conv.*, New York, 1997, preprint 4521.
[279] M. Vssnsnen, "Long term predictor for transform domain perceptual audio coding," in *Proc. 107th Audio Eng. Soc. Conv.*, Sept. 1999, preprint 5036.
[280] Y. Takamizawa, "An efficient tonal component coding algorithm for MPEG-2 audio NBC," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-97)*, 1997, pp. 331–334.
[281] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer Academic, 1992.
[282] T. Sreenivas and M. Dietz, "Vector quantization of scale factors in advanced audio coder (AAC)," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-98)*, May 1998.
[283] ——, "Improved AAC performance @ < 64 kb/s using VQ," in *Proc. 104th Audio Eng. Soc. Conv.*, Amsterdam, The Netherlands, 1998, preprint 4750.
[284] J. Herre and D. Schulz, "Extending the MPEG-4 AAC codec by perceptual noise substitution," in *Proc. 104th Audio Eng. Soc. Conv.*, Amsterdam, The Netherlands, 1998, preprint 4720.
[285] L. Contin, B. Edler, D. Meares, and P. Schreiner, "Tests on MPEG-4 audio codec proposals," *Signal Process. Image Commun. J.*, Oct. 1996.
[286] B. Edler, "Current status of the MPEG-4 audio verification model development," in *Proc. 101st Conv. Aud. Eng. Soc.*, Nov. 1996, preprint 4376.
[287] R. Koenen, F. Pereira, and L. Chiariglione, "MPEG-4: Context and objectives," *Signal Process. Image Commun. J.*, Oct. 1996.
[288] *Overview of the MPEG-4 standard*, Available WWW: http://www.cselt.it/mpeg/standards/mpeg-4/mpeg-4.html, July 1998.
[289] K. Brandenburg and M. Bosi, "Overview of MPEG audio: Current and future standards for low-bit-rate audio coding," *J. Audio Eng. Soc.*, pp. 4–21, Jan./Feb. 1997.
[290] S. Quackenbush, "Coding of natural audio in MPEG-4," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-98)*, May 1998.
[291] S. Park, Y. Kim, and Y. Seo, "Multi-layer bit-sliced bit-rate scalable audio coding," in *Proc. 103rd Conv. Aud. Eng. Soc.*, Sept. 1997, preprint 4520.
[292] E. Scheirer, "The MPEG-4 structured audio standard," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-98)*, May 1998.
[293] ——, "The MPEG-4 structured audio standard," in *Proc. ICASSP-98*, May 1998.
[294] E. Scheirer, "Structured audio and effects processing in the MPEG-4 multimedia standard," *ACM Multimedia Syst.*, vol. 7, no. 1, p. 11, 1999.
[295] E. Scheirer *et al.*, "AudioBIFS: The MPEG-4 standard for effects processing," in *Proc. DAFX98 Workshop on Digital Audio Effects Processing*, Nov. 1998.
[296] E. Scheirer, "The MPEG-4 structured audio orchestra language," in *Proc. ICMC*, Oct. 1998.
[297] E. Scheirer and L. Ray, "Algorithmic and wavetable synthesis in the MPEG-4 multimedia standard," in *Proc. 105th AES Conv.*, Sept. 1998, preprint 4811.
[298] B. Vercoe, W. Gardner , and E. Scheirer , "Structured audio: Creation, transmission, and rendering of parametric sound representations," *Proc. IEEE*, vol. 86, pp. 922–940, May 1998.
[299] *MPEG-4 structured audio homepage* [Online]. Available: HTTP: http://sound.media.mit.edu/~eds/~mpeg4
[300] A. Hoogendoorn, "Digital compact cassette," *Proc. IEEE*, vol. 82, pp. 1479–1489, Oct. 1994.
[301] T. Yoshida, "The rewritable MiniDisc system," *Proc. IEEE*, vol. 82, pp. 1492–1500, Oct. 1994.
[302] K. Tsutsui, "ATRAC (adaptive transform acoustic coding) and ATRAC 2," in *The Digital Signal Processing Handbook*, V. Madisetti and D. Williams, Eds. Boca Raton, FL: CRC Press, 1998, pp. 43.16–43.20.
[303] K. Tsutsui, H. Suzuki, O. Shimoyoshi, M. Sonohara, K. Akagiri, and R. Heddle, "ATRAC: adaptive transform acoustic coding for MiniDisc," *Collected Papers on Digital Audio Bit-Rate Reduction*, pp. 95–101, 1996.
[304] H. Yamauchi, K. Akagiri, M. Katakura, E. Saito, M. Kohut, M. Nishiguchi, and K. Tsutsui, "The SDDS system for digitizing film sound," in *The Digital Signal Processing Handbook*, V. Madisetti and D. Williams, Eds. Boca Raton, FL: CRC Press, 1998, pp. 43.6–43.12.
[305] J. Johnston and A. Ferreira, "Sum-difference stereo transform coding," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-92)*, May 1992, pp. II-569–II-572.
[306] D. Sinha, J. D. Johnson, S. Dorward, and S. Quackenbush, "The perceptual audio coder (PAC)," in *The Digital Signal Processing Handbook*, V. Madisetti and D. Williams, Eds. Boca Raton, FL: CRC Press, 1998, pp. 42.1–42.18.
[307] J. Johnston, D. Sinha, S. Dorward, and S. Quackenbush, "AT&T perceptual audio coding (PAC)," *Collected Papers on Digital Audio Bit-Rate Reduction*, pp. 73–81, 1996.

[308] D. Sinha and J. Johnston, "Audio compression at low bit rates using a signal adaptive switched filterbank," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-96)*, May 1996, pp. 1053–1056.
[309] B.C. J. Moore, *Introduction to the Psychology of Hearing*. New York: Academic, 1977.
[310] ISO-II, JTC1/SC29/WG11 N1420, "Report on the MPEG/audio multichannel formal subjective listening tests," ISO/MPEG-II Audio Committee, ISO/MPEG doc. MPEG94/063, 1994.
[311] D. Sinha and C. E. W. Sundberg, "Unequal error protection (UEP) for perceptual audio coders," in *Proc. 104th Aud. Eng. Soc. Conv.*, May 1998, preprint 4754.
[312] G. Davidson, L. Fielder, and M. Antill, "Low-complexity transform coder for satellite link applications," in *Proc. 89th Conv. Aud. Eng. Soc.*, 1990, preprint 2966.
[313] L. Fielder and G. Davidson, "AC-2: A family of low complexity transform-based music coders," in *Proc. 10th AES Int. Conf.*, Sept. 1991.
[314] G. Davidson and M. Bosi, "AC-2: High quality audio coding for broadcasting and storage," in *Proc. 46th Annu. Broadcast Eng. Conf.*, Apr. 1992, pp. 98–105.
[315] M. Davis, "The AC-3 multichannel coder," in *Proc. 95th Conv. Aud. Eng. Soc.*, Oct. 1993, preprint 3774.
[316] C. Todd, G. Davidson, M. Davis, L. Fielder, B. Link, and S. Vernon, "AC-3: Flexible perceptual coding for audio transmission and storage," in *Proc. 96th Conv. Aud. Eng. Soc.*, Feb. 1994, preprint 3796.
[317] G. Davidson, "Digital audio coding: Dolby AC-3," in *The Digital Signal Processing Handbook*, V. Madisetti and D. Williams, Eds. Boca Raton, FL: CRC Press, 1998, pp. 41.1–41.21.
[318] J. Blauert, *Spatial Hearing*. Cambridge, MA: MIT Press, 1974.
[319] M. Davis and C. Todd, "AC-3 operation, bitstream syntax, and features," in *Proc. 97th Conv. Aud. Eng. Soc.*, 1994, preprint 3910.
[320] S. Vernon, "Design and implementation of AC-3 coders," *IEEE Trans. Consumer Elec.*, Aug. 1995.
[321] S. Vernon, V. Fruchter, and S. Kusevitzky, "A single-chip DSP implementation of a high-quality, low bit-rate multi-channel audio coder," in *Proc. 95th Conv. Aud. Eng. Soc.*, 1993, preprint 3775.
[322] K. Terry and J. Seaver, "A real-time, multichannel dolby AC-3 audio encoder implementation," in *Proc. 101st Conv. Aud. Eng. Soc.*, Nov. 1996, preprint 4363.
[323] *Digital Television Standard, United States Advanced Television Systems Committee (ATSC). [Online]. Available: WWW: http://www.atsc.org/Standards/A53/*, Sept., 1995 Doc. A/53.
[324] Digital Audio-Visual Council (DAVIC), "DAVIC technical specification 1.2—Part 9," *Information Representation. [Online]. Available: WWW: http://www.davic.org*, Dec. 1996.
[325] T. Ryden, "Using listening tests to assess audio codecs," *Collected Papers on Digital Audio Bit-Rate Reduction*, pp. 115–125, 1996.
[326] S. Bech, "Selection and training of subjects for listening tests on sound-reproducing equipment," *J. Aud. Eng. Soc.*, pp. 590–610, July/Aug. 1992.
[327] International Telecommunications Union, "Subjective assessment of sound quality," Radio Communications Sector (ITU-R), Dusseldorf, Germany, CCIR Rec. 562-3, pt. 1, 1990.
[328] International Telecommunications Union, "Telephone transmission quality subjective opinion tests," Radio Communications Sector (ITU-R), Rec. P.80, 1994.
[329] T. Sporer, "Evaluating small impairments with the mean opinion scale—Reliable or just a guess?," in *Proc. 101st Conv. Aud. Eng. Soc.*, Nov. 1996, preprint 4396.
[330] M. Keyhl, C. Schmidner, T. Sporer, and R. Peterson, "Quality assurance tests of MPEG encoders for a digital broadcasting system—Part II: Minimizing subjective test efforts by perceptual measurements," in *Proc. 104th Conv. Aud. Eng. Soc.*, May 1998, preprint 4753.
[331] "Subjective performance assessment of telephone-band and wideband digital codecs," International Telecommunications Union, Dusseldorf, Germany, Rec. P.830, 1996.
[332] K. Precoda and T. Meng, "Listener differences in audio compression evaluations," *J. Aud. Eng. Soc.*, vol. 45, no. 9, pp. 708–715, Sept. 1997.
[333] S. Schiffman, M. Reynolds, and F. Young, *Introduction to Multidimensional Scaling: Theory, Method, and Applications*. New York: Academic, 1981.
[334] S. Shlien and G. Soulodre, "Measuring the characteristics of "expert" listeners," in *Proc. 101st Conv. Aud. Eng. Soc.*, Nov. 1996, preprint 4339.
[335] A. Milne, "New test methods for digital audio data compression algorihtms," in *Proc. 11th Int. Conf. Aud. Eng. Soc.*, May 1992, pp. 210–215.
[336] D. Moulton and M. Moulton, "Measurement of small impairments of perceptual audio coders using a 3-facet rasch model," in *Proc. 104th Conv. Aud. Eng. Soc.*, May 1998, preprint 4709.
[337] D. Moulton and M. Moulton, "Codec 'transparency,' listener 'severity,' program 'intolerance': Suggestive relationships between Rasch measures and some background variables," in *Proc. 105th Conv. Aud. Eng. Soc.*, Sept. 1998, preprint 4843.
[338] J. Linacre, *Many-Facet Rasch Measurement*. Chicago, IL: MESA Press, 1994.
[339] G. Rasch, *Probabilistic Models for Some Intelligence Attainment Tests*. Chicago, IL: Univ. of Chicago Press, 1980.
[340] M. Karjaleinen, "A new auditory model for the evaluation of sound quality of audio systems," in *Proc. ICASSP-85*, May 1985, pp. 608–611.
[341] K. Brandenburg, "Evaluation of quality for audio encoding at low bit rates," in *Proc. 82nd Conv. Aud. Eng. Soc.*, Mar. 1987, preprint 2433.
[342] J. Beerends and J. Stemerdink, "Measuring the quality of audio devices," in *Proc. 90th Conv. Aud. Eng. Soc.*, Feb. 1991, preprint 3070.
[343] ——, "A perceptual audio quality measure," in *Proc. 92nd Conv. Aud. Eng. Soc.*, Mar. 1992, preprint 3311.
[344] K. Brandenburg and T. Sporer, ""NMR" and "masking flag": Evaluation of quality using perceptual criteria," in *Proc. 11th Int. Conf. Aud. Eng. Soc.*, May 1992, pp. 169–179.
[345] B. Paillard, P. Mabilleau, and S. Morissette, "PERCEVAL: Perceptual evaluation of the quality of audio signals," *J. Aud. Eng. Soc.*, vol. 40, no. 1/2, pp. 21–31, Jan./Feb. 1992.
[346] J. Beerends and J. Stemerdink, "Modeling a cognitive aspect in the measurement of the quality of music codecs," in *Proc. 96th Conv. Aud. Eng. Soc.*, 1994, preprint 3800.
[347] ——, "Measuring the quality of speech and music codecs: An integrated psychoacoustic approach," in *Proc. 98th Conv. Aud. Eng. Soc.*, 1995, preprint 3945.
[348] J. Beerends, W. van den Brink, and B. Rodger, "The role of informational masking and perceptual streaming in the measurement of music codec quality," in *Proc. 100th Conv. Aud. Eng. Soc.*, May 1996, preprint 4176.
[349] T. Thiede and E. Kabot, "A new perceptual quality measure for bit rate reduced audio," in *Proc. 100th Conv. Aud. Eng. Soc.*, May 1996, preprint 4280.
[350] "Comparison between NMR, PERCEVAL, and PAQM as predictors of the subjective audio quality," International Telecommunications Union, Radio Communications Sector (ITU-R), Doc. 10-4/1-E, July 1994.
[351] T. Sporer, "Objective audio signal evaluation—Applied psychoacoustics for modeling the perceived quality of digital audio," in *Proc. 103rd Conv. Aud. Eng. Soc.*, Sept. 1997, preprint 4512.
[352] "Method for objective measurements of perceived audio quality," International Telecommunications Union, Radio Communications Sector (ITU-R), Rec. BS.1387.
[353] T. Painter and A. Spanias, *From G.722 to MP3 and Beyond: Algorithms for Perceptual Audio Coding*, monograph in final preparation.
[354] G. Soulodre *et al.*, "Subjective evaluation of state-of-the-art two-channel audio codecs," *J. Aud. Eng. Soc.*, vol. 46, no. 3, pp. 164–177, Mar. 1998.
[355] "Low bit rate audio coding," ITU-R, Geneva, Switzerland, Rec. BS.1115, 1997.
[356] U. Wustenhagen, B. Feiten, and W. Hoeg, "Subjective listening test of multichannel audio codecs," in *Proc. 105th Conv. Aud. Eng. Soc.*, Sept. 1998, preprint 4813.
[357] "Multi-channel stereophonic sound system with and without accompanying picture," ITU-R, Rec. BS.775-1, Nov. 1993.
[358] 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.*, Berlin, Germany, Mar. 1993, preprint 3458.
[359] R. K. Jurgen, "Broadcasting with digital audio," *IEEE Spectrum*, pp. 52–59, Mar. 1996.
[360] W. Pritchard and M. Ogata, "Satellite Direct Broadcast," *Proc. IEEE*, vol. 78, pp. 1116–1140, July 1990.
[361] P. Craven and M. Gerzon, "Lossless coding for audio discs," *J. Aud. Eng. Soc.*, pp. 706–720, Sept. 1996.

[362] *Digital Audio Compression Standard (AC-3)*, United States Advanced Television Systems Committee (ATSC), Doc. A/52, Dec 1995. [Online]. Available: WWW: http://www.atsc.org/Standards/A52/.
[363] C. Todd, G. Davidson, M. Davis, L. Fielder, B. Link, and S. Vernon, "AC-3: Flexible perceptual coding for audio transmission and storage," in *Proc. 96th Conv. Aud. Eng. Soc.*, Feb. 1994, preprint 3796.
[364] M. Dietz, H. Popp, and K. Brandenburg, "Audio compression for network transmission," *J. Audio Eng. Soc.*, pp. 58–70, Jan./Feb. 1996.
[365] T. Yoshida, "The rewritable MiniDisc system," *Proc. IEEE*, vol. 82, pp. 1492–1500, Oct. 1994.
[366] 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.
[367] A. Hoogendoorn, "Digital compact cassette," *Proc. IEEE*, vol. 82, pp. 1479–1489, Oct. 1994.
[368] ISO/IEC, JTC1/SC29/WG11 MPEG94/443, "Requirements for low bitrate audio coding/MPEG-4 audio,", MPEG-4, 1994.
[369] Y. Mahieux and J. P. Petit, "High-quality audio transform coding at 64 kb/s," *IEEE Trans. Commun.*, vol. 42, pp. 3010–3019, Nov. 1994.
[370] M. Purat and P. Noll, "Audio coding with a dynamic wavelet packet decomposition based on frequency-varying modulated lapped transforms," in *Proc. ICASSP-96*, May 1996, pp. 1021–1024.
[371] D. Sinha and A. Tewfik, "Low bit rate transparent audio compression using adapted wavelets," *IEEE Trans. Signal Processing*, vol. 41, pp. 3463–3479, Dec. 1993.
[372] J. Princen and J. D. Johnston, "Audio coding with signal adaptive filterbanks," in *Proc. ICASSP-95*, May 1995, pp. 3071–3074.
[373] D. Sinha and J. D. Johnston, "Audio compression at low bit rates using a signal adaptive switched filterbank," in *Proc. ICASSP-96*, May 1996, pp. 1053–1056.
[374] L. Mainard and M. Lever, "A bi-dimensional coding scheme applied to audio bitrate reduction," in *Proc. ICASSP-96*, May 1996, pp. 1017–1020.
[375] S. Boland and M. Deriche, "High quality audio coding using multipulse LPC and wavelet decomposition," in *Proc. ICASSP-95*, May 1995, pp. 3067–3070.
[376] P. Monta and S. Cheung, "Low rate audio coder with hierarchical filterbanks and lattice vector quantization," in *Proc. ICASSP-94*, May 1994, pp. II-209–II-212.
[377] D. Schulz, "Improving audio codecs by noise substitution," *J. Audio Eng. Soc.*, pp. 593–598, July\Aug. 1996.
[378] B. Grill, J. Herre, K. Brandenburg, E. Eberlein, J. Koller, and J. Muller, "Improved MPEG-2 audio multi-channel encoding," in *Proc. 96th Conv. Aud. Eng. Soc.*, Feb. 1994, preprint 3865.
[379] 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.*, May 1996, preprint 4196.
[380] E. Allamanche, R. Geiger, J. Herre, and T. Sporer, "MPEG-4 low delay audio coding based on the AAC codec," in *Proc. 106th Conv. Aud. Eng. Soc.*, May 1999, preprint 4929.
[381] K. Brandenburg and B. Grill, "First ideas on scalable audio coding," in *Proc. 97th Conv. Aud. Eng. Soc.*, Nov. 1994, preprint 3924.
[382] B. Grill and K. Brandenburg, "Two- or three-stage bit-rate scalable audio coding system," in *Proc. 99th Conv. Aud. Eng. Soc.*, Oct. 1995, preprint 4132.
[383] A. Spanias and T. Painter, "Universal speech and audio coding using a sinusoidal signal model," ASU-TRC, Jan. 1997.
[384] A. Jin, T. Moriya, T. Norimatsu, M. Tsushima, and T. Ishikawa, "Scalable audio coder based on quantizer units of MDCT coefficients," in *Proc. ICASSP-99*, Mar. 1999, pp. 897–900.
[385] J. Herre, E. Allamanche, K. Brandenburg, M. Dietz, B. Teichmann, and B. Grill, "The integrated filterbank based scalable MPEG-4 audio coder," in *Proc. 105th Conv. Aud. Eng. Soc.*, Sept. 1998, preprint 4810.
[386] M. Hans and R. Schafer, "An MPEG audio layered transcoder," in *Proc. 105th Conv. Aud. Eng. Soc.*, Sept. 1998, preprint 4812.
[387] B. Grill and B. Teichmann, "Scalable joint stereo coding," in *Proc. 105th Conv. Aud. Eng. Soc.*, Sept. 1998, preprint 4851.
[388] L. Ben and M. Sandler, "Joint source and channel coding for internet audio transmission," in *Proc. 106th Conv. Aud. Eng. Soc.*, May 1999, preprint 4932.
[389] S. Ramprashad, "A two-stage hybrid embedded speech/audio coding structure," in *Proc. ICASSP-98*, vol. I, May 1998, pp. 337–340.
[390] T. Moriya, N. Iwakami, A. Jin, K. Ikeda, and S. Miki, "A design of transform coder for both speech and audio signals at 1 bit/sample," in *Proc. ICASSP-97*, Apr. 1997, pp. 1371–1374.
[391] G. Schuller, "Time-varying filter banks with low delay for audio coding," in *Proc. 105th Conv. Aud. Eng. Soc.*, Sept. 1998, preprint 4809.
[392] F. Baumgarte, "Evaluation of a physiological ear model considering masking effects relevant to audio coding," in *Proc. 105th Conv. Aud. Eng. Soc.*, Sept. 1998, preprint 4789.
[393] Y. Huang and T. Chiueh, "A new forward masking model and its application to perceptual audio coding," in *Proc. ICASSP-99*, Mar. 1999, pp. 905–908.
[394] C. Lanciani and R. Schafer, "Subband-domain filtering of MPEG audio signals," in *Proc. ICASSP-99*, Mar. 1999, pp. 917–920.
[395] C. Neubauer and J. Herre, "Digital watermarking and its influence on audio quality," in *Proc. 105th Conv. Aud. Eng. Soc.*, Sept. 1998, preprint 4823.
[396] A. Tewfik, M. Swanson, and B. Zhu, "Data embedding in audio: Where do we stand," in *Proc. ICASSP-99*, Mar. 1999, p. 2075.
[397] *"Method for objective measurements of perceived audio quality,", ITU-R BS.1387, 1998.
[398] K. Konstantinides, "Fast subband filtering in MPEG audio coding," *IEEE Signal Processing Lett.*, vol. 1, pp. 26–28, Feb. 1994.
[399] T. Trinkaus, "Perceptual coding of audio and diverse speech signals," Ph.D. dissertation, Georgia Institute of Technology, Atlanta, Dec. 1999.
[400] T. Trinkaus and M. Clements, "An algorithm for compression of wideband diverse speech and audio signals," in *Proc. ICASSP-99*, Mar. 1999, pp. 901–904.
[401] R. Arean, J. Kovacevic, and V. Goyal, "Multiple description perceptual audio coding with correlating transforms," *IEEE Trans. Speech Audio*, vol. 8, pp. 140–145, Mar. 2000.
[402] T. Thiede, W. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, "PEAQ-The ITU standard for objective measurement of perceived audio quality," *J. Aud. Eng. Soc*, vol. 48, pp. 3–29, Jan./Feb. 2000.
[403] W. Treurniet and G. Soulodre, "Evaluation of the ITU-R objective audio quality measurement method," *J. Aud. Eng. Soc*, pp. 164–173, Mar. 2000.
[404] G. Schuller and T. Karp, "Modulated filter banks with arbitrary system delay: Efficient implementations and the time-varying case," *IEEE Trans. Signal Processing*, vol. 48, pp. 737–748, Mar. 2000.
[405] J. O. Smith, III and J. Abel, "Bark and ERB bilinear transforms," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 697–708, Nov. 1999.



Ted Painter (S'95) was born in Boston, MA, in 1967. He received the A.B. degree in engineering sciences and computer science from Dartmouth College, Hanover, NH, in 1989 and the M.S. degree in electrical engineering from Arizona State University (ASU), Tempe, in 1995. He is currently pursuing the Ph.D. degree in electrical engineering at the ASU Telecommunications Research Center (TRC).

Prior to joining ASU, he was an Embedded Systems Development Engineer with Applied Systems, Gilbert, AZ, from 1989 to 1992, and then an Industrial Fellow with the Flight Controls Group at Honeywell Commercial Flight Systems, Phoenix, AZ, from 1992 to 1994. He has been a Research Associate with the ASU TRC since 1995. He recently joined the Technical Staff of the StrongARM Systems Engineering Group at Intel Corp., Hudson, MA. His primary research interests are in the areas of speech and audio signal processing, perceptual coding, and psychoacoustics.

Mr. Painter is a student member of the Audio Engineering Society.

Andreas Spanias (S'84–M'85–SM'94), is a Professor in the Department of Electrical Engineering at Arizona State University (ASU), Tempe. His research interests are in the areas of adaptive signal processing and speech processing. He has been Principal Investigator on research contracts from Intel Corporation, Sandia National Labs, Motorola Inc., and Active Noise and Vibration Technologies. He has also consulted with Inter-Tel Communications, Texas Instruments, and the Cyprus Institute of Neurology and Genetics.

He is a member of the DSP Committee of the IEEE Circuits and Systems Society, and has served as a member in the Technical Committee on Statistical Signal and Array Processing of the IEEE Signal Processing Society. He has also served as Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and as General Co-chair of the 1999 International Conference on Acoustics Speech and Signal Processing (ICASSP-99) in Phoenix, AZ. He is currently the IEEE Signal Processing Vice-President for Conferences and the Chair of the Signal Procesing Conference Board. He is also member of the IEEE Signal Processing Executive Committee and Associate Editor of the IEEE SIGNAL PROCESSING LETTERS.