Signal Compression Based on Models of Human Perception

NIKIL JAYANT, FELLOW, IEEE, JAMES JOHNSTON, AND ROBERT SAFRANEK, SENIOR MEMBER, IEEE

The problem of signal compression is to achieve a low bit rate in the digital representation of an input signal with minimum perceived loss of signal quality. In compressing signals such as speech, audio, image, and video, the ultimate criterion of signal quality is usually that judged or measured by the human receiver. As we seek lower bit rates in the digital representations of these signals, it is imperative that we design the compression (or coding) algorithm to minimize perceptually meaningful measures of signal distortion, rather than more traditional and tractable criteria such as the mean squared difference between the waveforms at the input and output of the coding system.

This paper develops the notion of ***perceptual coding*** *based on the concept of distortion masking by the signal being compressed, and describes how the field has progressed as a result of advances in classical coding theory, modeling of human perception, and digital signal processing. We propose that fundamental limits in the science can be expressed by the semi-quantitative concepts of* ***perceptual entropy*** *and the* ***perceptual distortion-rate function****, and we examine current compression technology with respect to that framework. We conclude with a summary of future challenges and research directions.*

I. Introduction

The problem of signal compression is to achieve a low bit rate in the digital representation of an input signal with minimum perceived loss of signal quality. The function of *compression* is often referred to as *low bit rate coding*, or *coding*, for short. As we seek lower bit rates in the digital representation of a signal, it is imperative that we design the coding (or compression) algorithm to minimize a perceptually meaningful measure of signal distortion, rather than more traditional and more tractable criteria such as the mean squared difference between the waveforms at the input and output of the coding system.

Central to the above idea is the notion of distortion masking or noise masking, whereby the distortion (or noise) that is inevitably introduced in the coding process, *if properly distributed or shaped*, is masked by the input signal itself. The masking can be partial or total, leading either to increased quality compared to a system without noise shaping, or to perfect signal quality that is equivalent to that of the uncoded signal. In either case, the masking occurs because of the inability of the human perceptual mechanism to distinguish two signal components (one belonging to the signal, one belonging to the noise) in the same spectral, temporal, or spatial locality. An important effect of this limitation is that the perceptibility of noise can be zero even if the objectively measured local signal-to-noise ratio is modest or low. If the noise distribution is such that the masking effect is successfully invoked at all points in frequency, time, and space, the result will be a global coding operation of high, or perfect signal quality.

Ideally, the noise level at all points in the signal space is exactly at the level of *just-noticeable distortion* (JND). This corresponds to perfect signal quality at the lowest possible bit rate. The signal is now neither overcoded nor undercoded. This bit rate is a fundamental limit to which we can compress the signal with zero (perceived) distortion. We call this limit the *perceptual entropy*. A signal coding algorithm that is based on the criterion of minimizing the perceived error is called a *perceptual coding* algorithm.

The basic notion of maximizing perceived signal quality rather than the more tractable mean squared error is not novel. However, significant progress is now being made in the sophistication, degree, and dynamics of perceptual coding. Models of human audition and vision, while still imperfect, are leading to more accurate models for JND and noise masking. In parallel, the arithmetic capabilities of digital signal processors have increased to the point where the computational complexity of perceptual coding can be supported in practical hardware.

While perceptual coding is important for all the signals mentioned earlier, it is most significant for signal classes that lack a good source model. Speech signals do have a universal and reliable production model, based on our knowledge of the human vocal apparatus. The source model leads directly to efficient models for removing signal redundancy. These models are utilized in techniques such as linear predictive coding (LPC) to reduce the bit rate while maintaining a specified level of signal quality. In general,

Manuscript received November 2, 1992; revised July 7, 1993.

The authors are with the Signal Processing Research Department , AT&T Bell Laboratories, Murray Hill, NJ 07974.

IEEE Log Number 9212430.

0018-9219/93$03.00 © 1993 IEEE

audio, image, and video signals do not have a similar single or universal *source* model. As a result, more of the burden for bit rate reduction falls on the *sink* model—the notion of shaping the noise into components of distortion that are perceptually unnoticed, the paradigm of perceptual coding.

The topics of speech, audio, and image coding constitute significant separate topics in their own right. The purpose of this paper, however, is to summarize generic tools of perceptual coding, while drawing freely from signal-specific implementations in individual sections of the paper. In order to cover this broad ground in a single article, we have assumed a basic knowledge of the principles of coding [65], and we have taken the liberty of writing subsections that are sometimes shallow and cryptic.

The promises of perceptual coding are significant. But several issues need to be addressed in order to achieve its maximum potential. Traditional perceptual models have been generally based on global knowledge. As a result, they are not locally optimal, but on the other hand are fairly robust across variations such as those in input signal or viewing distance (in the case of image and video coding). The new generation of signal-dependent or dynamic models which are based on local properties are more powerful in the context of input nonstationarity, but they can also be more sensitive to the accuracy of signal analysis, and to the correctness of the mapping from the signal analysis to the JND model for quantization. Designing dynamic perceptual systems that combine peak performance with robustness is a formidable research problem.

The sections of the paper are organized as follows.

Section II is an essay on signal compression. This includes a description of four dimensions of performance: *signal quality, bit rate, algorithm complexity*, and *communication delay*. It is followed by a summary of applications, standards and technology goals. Section II also discusses two important tools in the trade: reduction of redundancy in the input signal, and the removal of irrelevant information in the operation of quantization. Focus in this paper will be on the second operation, the utilization of *perceptual irrelevancy* for realizing high quality at low bit rates.

Section III provides a very brief and pragmatic discussion of the properties of audiovisual signals, and the physics of the human perceptual mechanism, as utilized in our examples of perceptual coding.

Section IV describes how certain basic building blocks in signal compression algorithms are qualitatively matched to the properties and methods of human perception. These building blocks, filter banks for example, provide an advantageous framework for utilizing signal redundancy as well as for the subsequent utilization of the irrelevancy principle by the quantizing system.

Section V discusses amplitude quantization. This is the part of the compression algorithm where the flexibilities afforded by the perceptual mechanism are utilized in conserving the bit rate in the digital representation of the signal being coded. This section of the paper describes the evolution of perceptual coding. It begins with primitive examples of quantizers that minimize a criterion more useful than minimum mean squared error (mmse), and it concludes with systems driven by the perceptual notions of *just-noticeable distortion* (JND) and *minimally noticeable distortion* (MND).

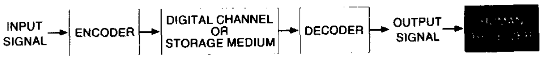


Fig. 1. Digital coding for signal compression.

Sections VI–IX discuss perceptual coding algorithms that have proven useful in the compression of speech, audio, image, and video signals, and point out the extent to which these algorithms have impacted current technology and international standards for signal compression.

Section X is a summary of research directions. It describes the need to incorporate notions of perceptual coding even more strongly in existing standards as well as in evolving technology. It also describes research advances that will be needed to make perceptual coding less of an art, and to establish it as a robust and well-understood scientific discipline.

II. Overview of Signal Compression

The material in this section is borrowed from a recent article by one of the authors [62]. The first two sections of that article are reproduced here as background for the subsequent discussion of perceptual coding in the remainder of this paper. Readers well-versed in the theory of signal compression and its application to speech, audio, and image signals may choose to skip this section and proceed directly to Section III.

A. Signal Compression

Signal coding is the process of representing an information signal in a way that realizes a desired communications objective such as analog-to-digital conversion, low-bit-rate transmission, or message encryption. In the literature, the terms *source coding*, *digital coding*, *data compression*, *bandwidth compression*, and *signal compression* are all used to connote the function of achieving a compact digital representation of a signal, including the important subclass of analog signals such as speech, audio, and image. When the terms *coding*, *encoding*, and *decoding* are used in this paper, they will all refer to the specific common objective of compression. An important theme of our discussion is the human receiver at the end of the communication process (Fig.1).

The characteristics of human perception are of relevance to the optimization of a variety of functions in digital communication networks [170]. For example, one needs to consider the subjective consequences of communication delay [75], [79], [124], [160], differential delay, and jitter, as in some packet network designs [164], the effects of spatial disparity between audio and visual channels [81], and the effects of imperfect transmission channels [26], [134]. Focus in this paper is on the subjective effects of signal distortion introduced in the process of compression;

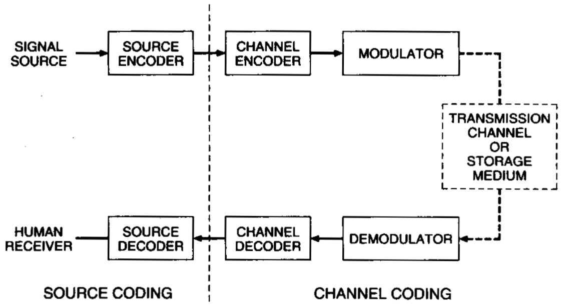


Fig. 2. Block diagram of a digital communication system.

and on the use of perceptual criteria as an integral part of coder design.

Figure 2 defines the role of signal compression (source coding) in digital communication. While the source coder attempts to minimize the necessary bit rate for faithfully representing the input signal, the *mo*dulator–*dem*odulator (modem) seeks to maximize the bit rate that can be supported in a given channel or storage medium without causing an unacceptable level p_e of bit error probability. The bit rate in source coding is measured in bits per sample or bits per second (bps). In modulation, the rate is measured in bits per second per Hertz (bps/Hz). The channel coding boxes add redundancy to the encoder bit stream for the purpose of error protection. In so-called coded modulation systems, the operations of channel coding and modulation are combined for greater overall efficiency. The processes of source and channel coding, as well as the process of multi-user networking, can sometimes be integrated to increase the efficiency of digital communication.

The capability of signal compression has been central to the technologies of robust long-distance communication, high-quality signal storage, and message encryption. Compression continues to be a key technology in communications in spite of the promise of optical transmission media of relatively unlimited bandwidth. This is because of our continued and, in fact, increasing need to use bandlimited media such as radio and satellite links, and bit-rate-limited storage media such as CD-ROM's and solid-state memory chips.

B. Background

The information-theoretical foundations of signal compression date back to the seminal work of Shannon [137], [138]. His mathematical exposition defined the information content or *entropy* of a source and showed that the source could be coded with zero error if the encoder used a transmission rate equal to or greater than the entropy, and if the encoder used a long processing delay, tending in general to infinity. In the special case of the infinite alphabet or analog source, the encoding error tends to approach zero only at an infinite bit rate. However, in practice, the error is close enough to zero at finite rates. In the case of a finite-alphabet or discrete-amplitude source, the entropy is finite, and the bit rate needed for zero encoding error is finite as well. An important example of a finite-entropy source is an analog signal stored in a computer as a sequence of discrete amplitudes. The raw (uncompressed) bit rate of such a signal is typically 8, 16, or 24 b/sample, respectively, for a grey-level image, high-quality speech (or audio), and a color image with three 8-b components. The entropy, or the minimum bit rate for zero encoding error, will be typically smaller because of the statistical redundancy in the input sequence.

The inadequacies of the classical source coding theory are twofold. First, the theory is nonconstructive, offering bounds on distortion-rate performance rather than techniques for achieving these targets. However, the classical theory teaches us important qualitative recipes such as delayed encoding, as in vector quantization with a large vector dimension or block length. Second, the source model used in the classical theory does not capture what are now recognized as fundamental nuances in audio and visual signal processing. These include the fact that the input signal is non-Gaussian, nonstationary, and in general has a complex and intractable power spectrum; and further, the observation that the human receiver does not employ a mean-squared-error criterion in judging the similarity of a coded signal to the uncoded original. As a result of the above complications, some of the observations of classical source coding do not carry over in an obvious way to signal compression as discussed in this paper. One such result is that the source entropy measured with a perceptual distortion criterion is different from, and generally much lower than, the classical entropy measured with a mean-squared-error criterion for coding distortion. Another classical result that needs to be re-examined is the thesis that in principle, the processes of source and channel coding can be separated without loss of optimality. This model would hold very nicely for the digital communication of data sequences, but does not necessarily suggest an optimal solution to the complex problem of communicating audio or visual signals over a noisy channel with high perceptual fidelity, robustness, and bit-rate efficiency.

The technology and literature of signal compression has therefore evolved independently, with basic and valuable inspiration from Shannon's theory [137], [138] and the rate-distortion theory that followed it [9], [33], [41], [58], [172], but also with a great deal of innovative engineering on the part of the practitioners of signal compression [1], [28], [42], [43], [49], [52], [57], [63], [65], [87], [93], [140], [141], speech coding [5], [6], [27], [36], [37], [44], [46], [54], [83], [90], [94], [122], [135], [148], [173], audio coding [11], [67], [96], [146], and image (and video) coding [29], [35], [39], [47], [48], [53], [84], [99], [100], [101], [103], [110], [118], [126], [131], [133], [168], [171]. In particular, work on speech compression has benefited greatly from studies of speech production and speech perception by humans, and research on visual perception has similarly impacted the parallel field of image compression. Although the mathematical theory of source coding is a common denominator, the fields of speech and image coding have been generally discussed by different schools, with a few recent exceptions. One of the purposes of this paper is to point out that as we address future technology targets in the disciplines of speech and image compression, common threads continue to exist. One such commonality is the increasing importance of matching the compression algorithm to the human perceptual mechanism—the auditory process in one case, and the visual process in the other. Another common element is the body of newly emerging techniques for quantization and time–frequency analysis, in support of perceptually tuned coding.

C. The Dimensions of Performance in Signal Compression

The generic problem in signal compression is to minimize the bit rate in the digital representation of the signal while maintaining required levels of signal quality, complexity of implementation, and communication delay. We will now provide brief descriptions of the above parameters of performance.

1) Signal Quality: Perceived signal quality is often measured on a five-point scale that is well known as the *mean opinion score* or *mos* scale in speech quality testing—an average over a large number of speech inputs, speakers, and listeners evaluating signal quality [32], [50], [65], [74], [76], [158]. The five points of quality are associated with a set of standardized adjectival descriptions: *bad, poor, fair, good,* and *excellent.* Every example of an input being evaluated is assigned one of these levels in the course of a subjective test. Five-point scales of quality have also been used in audio and image testing [96], [165]. An alternative methodology for subjective testing, better suited to image and audio work, is based on an inverted scale that categorizes levels of *impairment* [14], [15], [147] (*very annoying, annoying, slightly annoying, perceptible but not annoying*, and *imperceptible*). Other variations include the notion of averaging measurements over a selected *difficult* subset of input signals [45], [145], in order to provide conservative scores of coder performance. In some illustrations of this paper, we use the original notion stated in this paragraph, a quality (rather than an impairment) scale, and an average over a large set of typical inputs, in each of the four generic categories: telephone-bandwidth speech, CD-grade audio, still images, and video. Our quantitative discussion of image and video quality will be impressionistic at best, given the multidimensionality of the problem (dependence of subjective quality on input scene, picture resolution, image size, and viewing distance), and given the general lack of formal quality assessments in recent image coding literature. In the field of speech coding, *mos* evaluations are well-accepted and sometimes supplemented with measurements of *intelligibility*[155] and *acceptability* [154].

2) Bit Rate: We measure the bit rate of the digital representation in *bits per sample, bits per pixel* (bpp), or *bits per second* (bps), depending on context, where *pixel* (sometimes shortened to *pel*) refers to a *pi*cture *el*ement, or an image sample. The rate in bits per second is merely the product of the sampling rate and the number of bits per sample. The sampling rate is typically slightly higher than about twice the respectivesignal bandwidth, as required by the Nyquist sampling theorem [65]. Table 1 defines four commonly used grades of audio bandwidth. Typical sampling rates are 8 kHz for telephone speech, 16 kHz for AM-radio-grade audio, and 44.1 or 48 kHz for CD (compact-disk) audio or DAT (digital audio tape) audio, both of which are signals of 20-kHz bandwidth.

Table 1 Digital Audio Formats

Format	Sampling Rate (kHz)	Bandwidth (kHz)	Frequency Band
Telephony	8	3.2	(200–3400 Hz)
Teleconferencing	16	7.0	(50–7000 Hz)
Compact Disk (CD)	44.1	20.0	(20–20 000 Hz)
Digital Audio Tape (DAT)	48	20.0	(20–20 000 Hz)

Table 2 Digital Television Formats (CIF: Common Intermediate Format; CCIR: International Consultative Committee for Radio; HDTV: One Example of a High Definition Television Format)

Format	Spatio-Temporal Resolution	Sampling Rate
CIF	360 × 288 × 30	3 Mpps
CCIR	720 × 576 × 30	12 Mpps
HDTV	1280 × 720 × 60	60 Mpps

Table 2 defines commonly used grades of video in terms of sampling rate in pixels per second (pps) or Hertz (Hz). The sampling rates for the CIF, CCIR, and HDTV formats defined in the table are 3, 12, and 60 MHz. Respective Nyquist bandwidths are approximately 1.5, 6, and 30 MHz, although bandwidth-limiting of image and video signals is in general less formal than the bandlimiting operations used for speech and audio signals. The HDTV format in the table is merely a specific example, one of several alternative current formats. The sampling rates in the table refer to luminance information. Overheads for including color information are system-dependent. In the CIF format, the color overhead is 50% in sampling rate, corresponding to a 50% subsampling relative to luminance in each of the horizontal and vertical directions, and a total of two chrominance components. Higher degrees of subsampling are sometimes used, leading to overall color overheads lower than 50%. In the line-interlaced CCIR format, the subsampling of color is performed only in the horizontal direction, and the final overhead in sampling rate due to color is 100%.

3) Complexity: The complexity of a coding algorithm is the computational effort required to implement the encoding and decoding processes in signal processing hardware, and it is typically measured in terms of arithmetic capability and memory requirement. Coding algorithms of significant complexity are currently being implemented in real time, some of them on single-chip processors. Other, related measures of coding complexity are the physical size of the encoder, decoder, or codec (*enco*der plus *dec*oder), their cost (in dollars), and the power consumption (in watts or milliwatts, mW), a particularly important criterion for portable systems [24], [142].

The technology of digital signal processing (DSP) has been evolving quite rapidly. One aspect of this progress can be measured in terms of the millions of instructions per second (mips) that can be accommodated on a single general-purpose processor, as a function of time. The evolution is exponential, with no evidence of saturation in the near-term [89]. In the five-year period from 1990 to 1995, the typical per-chip capability is expected to increase tenfold, from about 25 mips to about 250 mips per chip. Supporting this evolution in arithmetic capability is a parallel advance in memory capability. The significance of the above advances is that sophisticated compression algorithms that demand increasing levels of complexity will be supported by DSP technology, in the form of single-chip, multichannel implementations of the relatively less complex algorithms, and realistic parallel-processing machines for the more complex techniques. Power dissipation and cost are also expected to decrease steadily, making DSP technology increasingly useful for personal portable devices and for other high-volume consumer applications.

4) Communication Delay: Increasing complexity in a coding algorithm is usually associated with increased processing delays in the encoder and decoder. Although improved DSP capability can be used as an argument in favor of more sophisticated algorithms, the need to constrain communication delay should not be underemphasized. This need places important practical restrictions on the permissible sophistication of a signal compression algorithm. Depending on the communication environment, the permissible total delay for one-way communication (coding plus decoding delay) can be as low as about 1 ms (as in network telephony under conditions of no echo control), and as high as about 500 ms (as in very low bit rate videotelephony (or *videophony*) where the delay performance is severely compromised in the interest of obtaining a received picture good enough for communication). In one-way applications such as broadcasting, processing delay is somewhat less relevant although one attempts to minimize it because of its impact on latency in station switching. Communication delay is largely irrelevant for one-way communication for storage, message-forwarding, and voice mail.

D. Coding and Digital Communication

Figure 3 describes performance criteria in digital communication, by recapitulating the four dimensions of coder performance, explained specifically for source coding in Section II-C. These dimensions of performance apply to channel coding and modulation as well, although the units of quality and bit rate are different. Respective units, defined either in Section II-C or in the third paragraph of Section II-A, appear along the *quality* and *efficiency* axes in Fig. 3. Along each axis, the left and right entries refer, respectively, to source and channel coding. The units of *complexity* and *delay* are identical for source and channel coding, although those parameters are used for different reasons in the two cases. Processing delay is used in source coding to remove signal redundancy or to obtain a framework for reducing irrelevancy in quantization. In channel coding, delay may be used for adding error protection bits and for processes such as interleaving for the randomization of burst errors.

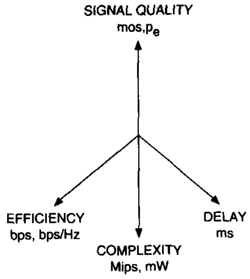


Fig. 3. The dimensions of coder performance.

The axes in Fig. 3 define a four-dimensional space in which some regions are theoretically allowable, and some regions are desirable for specific communication applications. Researchers in source and channel coding attempt to describe the allowable regions and the tradeoffs as quantitatively as possible. The focus of this paper is on the domain of source coding. In particular, we shall comment extensively on the quantitative relationship between compressed signal quality and bit rate.

E. Applications of Signal Compression

Figure 4 depicts various applications of signal compression. The vertical axis in the figure does not have any special meaning. The numbers on the horizontal axis are bit rates *after* compression. The labels in the figure represent, in an approximate fashion, current capabilities. The bit rates spanned by these labels, and in some cases, the bit rates on which the labels are centered, represent the rates at which compressed signals render the corresponding application practical. As our capabilities in compression improve, the labels in the figure tend to shift to the left. Signals covered in Fig. 4 include telephone speech, wideband audio (speech and music), and a wide range of image signals, including still pictures and motion video. In the following paragraphs, we provide very brief summaries of current capabilities in the compression of speech, audio, and image signals.

1) Telephone Speech: Speech compressed to 2.4 kbps provides a high level of intelligibility. However, speech quality, naturalness, and speaker recognizability are all poor at this bit rate. The need for digital encryption over a very wide range of transmission media is the main reason why 2.4-kbps speech is widely used, particularly in government and defense communications [148].

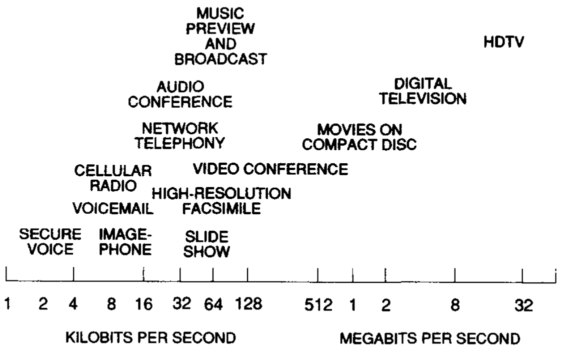


Fig. 4. Applications of signal compression.

A bit rate of 4.8 kbps is sufficient to provide measurable gains in naturalness and speaker recognition. This bit rate is also of interest to government and defense applications [72]. With the increased demands of mobile telephony over bandlimited channels, 4.8-kbps speech coding is also becoming very important for commercial communications using digital cellular radio. At 8 kbps, which is the bit rate chosen for first-generation digital cellular telephony in North America [44], speech quality is high, although significantly lower than that of the uncoded telephone-band signal.

At 16 kbps and beyond, speech quality is extremely close to that of the original, especially after a single stage of encoding and decoding. We use the term *network quality* to signify a performance level at which there is sufficient margin for additional functions such as multiple stages of encoding and decoding for speech, as well as high-accuracy transmission of nonspeech voiceband signals such as modem waveforms. The bit rate required for network-quality telephony, which was historically 64 kbps, and later reduced to 32 kbps [65], has now come down to 16 kbps [20].

The application of *voicemail* involves speech storage as well as speech transmission for forwarding the voice message. Depending on the network environment used for this service, and the desired speech quality in the received message, the bit rate can range from 4 to 32 kbps, with increased focus expected in the 4- to 16-kbps range in the future.

2) Wideband Speech: The 7-kHz speech signal has a higher voice quality than traditional telephony. This is partly due to increases in speaker presence and the naturalness of speech, as provided by the low-frequency enhancement (the added band from 200 to 50 Hz; see Table 1), and partly due to increased intelligibility and crispness provided by high-frequency enhancement (the added band from 3400 to 7000 Hz). The higher quality of wideband speech is desirable for the extended communication task of a long audioconference call. It is also appropriate for other applications of loudspeaker telephony and for systems that include a high-quality speakerphone. It is also known that inexpensive electret microphones can support an incoming bandwidth of 7 kHz.

The standardized bit rate for high-quality coding of 7-kHz speech is 64 kbps [16], [96], typically for an audioconference application using the Integrated Services Digital Network (ISDN). Recent algorithms have provided 7-kHz capability at 32 kbps [64], permitting stereo-teleconferencing or dual-language programming over basic rate ISDN. The projected capability for high-quality coding of 7-kHz speech is at least as low as 16 kbps [40], [127], [166].

The lower bit rates for wideband speech are also central to high-quality conferencing with combined audio and video. Current practice, at low values of total bit rate, such as 64 kbps, is to limit speech to the traditional telephone bandwidth of 3.2 kHz, and to use 8 kbps, or at the most, 16 kbps, for the coding of the audio channel. With advances in wideband speech compression, 16-kbps coding of 7-kHz audio can be expected to be an important component of audiovisual conferencing.

3) Wideband Audio: On a compact disk (CD), 20-kHz audio is sampled at 44.1 kHz and stored at 16 b/sample, or 706 kbps per sound channel. Current algorithms for audio compression [11], [45], [67], [98], [108], [145], [146] provide CD quality at 128 kbps per channel, and even at 64 kbps per channel for many audio inputs. These capabilities are important for emerging digital systems for audio broadcast and music preview. The capabilities are also central to applications that combine audio and visual functions, such as CD-ROM multimedia with a total bit rate of about 1.5 Mbps and digital television (Fig. 4) with multiple sound channels.

4) Still Images: A 500 × 500 pixel color image, with the uncompressed format of 24 bits per pixel (bpp), will require about 100 s of transmission time over a 64-kbps link. With 0.25-bpp coding, the transmission time is about 1 s, a number that would be deemed excellent for an interactive "slide show" [121]. Current technology for coding a 500 × 500 image is capable of providing good picture quality at 0.25 bpp for a wide class of color images, assuming a viewing distance of about 6 times the picture height [128], [129], [156]. For most images, increasing the bit rate to 1 bpp provides excellent and, in some cases, perceptually perfect image quality. The corresponding transmission time over a 64-kbps link is 4 s. High-resolution facsimile typically takes several seconds of transmission time over a 64-kbps channel even after the use of powerful techniques for fax compression [59], [62], [104], [112].

Techniques for *progressive transmission* [59], [156], [157] involve a first stage of coding characterized by a low bit rate and rapid picture access, followed if needed, by additional stages of transmission that upgrade the picture quality. Progressive transmission is ideal for applications such as *telebrowsing*. It is also appropriate for applications where one expects display modalities (terminals and printers) of varying resolutions. The price paid for this flexibility is that progressive systems provide a slightly lower signal quality at a given bit rate compared to a single-bit rate algorithm tuned to that specific rate.

The image-phone application in Fig. 4, which assumes the use of a telephone line and a 9.6-kbps modem, involves pictures of very low spatial and temporal resolution. A typical resolution would be 100 × 100 pixels per frame, and about 3 to 6 frames per second. With an even lower temporal resolution such as about 1 frame per second, the system degenerates to a sequence-of-stills service, sometimes referred to as *freeze-frame video.*

5) Digital Video: Comfortable videoconferencing requires CIF resolution (360 × 288 pixels per frame, Table 2), or at least quarter-CIF resolution (180 × 144 pixels per frame). Input temporal resolutions are usually submultiples of 30 frames per second, say 15 or even 10, for bit rates lower than about 1.5 Mbps. With CIF resolution and a bit rate of 1.5 Mbps, the communications quality of the service is generally agreed to be high. With quarter-CIF or somewhat lower resolution, and with correspondingly lower values of temporal resolution, it is possible to achieve lower bit rates such as 48 or 112 kbps. But the video quality is useful only if one accepts low levels of sharpness in the output picture, and very low levels of motion activity in the input scene, as in the head-and-shoulders view of a single person—an environment sometimes referred to as video*telephony*, rather than video*conferencing*. The bit rates of 48 and 112 kbps are appropriate for ISDN systems with total bit rates of 64 and 128 kbps, respectively, and a bit rate of 16 kbps for voice transmission. The bit rate of 384 kbps is an interesting number in the current state of the technology. At this bit rate, it is possible to provide a fair, if not high, level of picture quality in the coding of a videoconference scene.

CD-ROM media have a net throughput rate of about 1.5 Mbps for source data and a total bit capacity of a few gigabits. If video can be compressed to about 1 Mbps, a CD-ROM device could store and play out about an hour or more of the video signal, together with compressed stereo sound. This capability is central to various emerging applications of CD-ROM multimedia, including the specific example of a movie on an audio compact disk [85]. The additional capability of selecting a still-image snapshot of a desired part of the image sequence leads to the concept of *addressable video.* This is an important feature in emerging systems for video storage.

Uncompressed high-definition television (HDTV) has a bit rate of over a gigabit per second (the product of a sampling rate on the order of 60 MHz, as in Table 2, and the representation of three color components with a total of 24 b/sample). Compression of the HDTV signal to a bit rate on the order of a few tens of Mbps will create several important opportunities for HDTV broadcasting. In particular, a bit rate in the range of 20 Mbps will bring the service into the realm of a 6-MHz transmission channel [102], implying the capabilityof simulcasting the HDTV version of a program in vacant slots of an NTSC channel set. Transmission rates higher than 20 Mbps are appropriate for higher quality transmissions over satellite and broadband ISDN channels, cable transmission, and applications of HDTV for movie production. Rates of 2 to 8 Mbps are appropriate for digital television with CCIR resolution.

F. Compression Standards

The need for interoperating different realizations of signal encoding devices (transmitters) and signal decoding devices (receivers) has led to the formulation of several international and national standards for compression algorithms. The overview article [62] provides a nonexhaustive summary of compression standards for speech [20], [44], [65], [72], [148], audio [96], [98], [108], image [59], [156], and video [85], [88], [107]. Additional information provided in that article includes typical applications, typical levels of signal quality, and the approximate date of formulation of the standard.

The recent explosion in standards activity has had an important impact on research and development in the field. Standards have led to increased focus in applied research. They have sometimes stimulated highly productive new research as well. But they have also elevated the threshold of performance that a novel research algorithm needs to exceed before it is widely accepted, given that the supplanting of a recently endorsed standard is generally difficult and expensive.

Several applications of signal compression are decoder-intensive in the sense that users need access only to a decoder, the encoding being a one-time operation by the provider of the service. Examples are multimedia, Digital Audio Broadcast (DAB), and HDTV decoders. In recognition of this, corresponding standards have specified the decoder algorithm and bit-stream syntax rather than the encoder. In these cases, compatible enhancements to the standard are possible in the encoding module, as well as in optional modules of pre- and postprocessing: pre-filtering at the encoder, and post-filtering at the decoder.

G. Quality of the Compressed Signal

We have noted earlier that there are several dimensions defining the performance of a coding system. If we ignore the dimensions of algorithmic complexity and communication delay for the moment, coder improvements can be demonstrated in two ways: by measuring signal quality improvement at a specified bit rate, or by realizing a specified level of signal quality at a lower bit rate. Depending on the application, one of the above approaches would be more relevant than the other. For example, in the problems of coding telephone speech at 4 kbps and HDTV at 15 to 20 Mbps, the bit rates are defined by important generic applications, and the goal of coding research is to enhance signal quality at those rates. On the other hand, in the field of digital audio broadcasting (DAB), where the signal quality needs to be transparent to the coding algorithm and equivalent, say, to that of CD audio, the goal is to demonstrate such performance at progressively lower rates (say, at an average of 48 or 64 kbps per channel [69], rather than at 96 or 128 kbps per channel as in currently popular algorithms).

So far, our discussions of bit rate have been in terms of kilobits per second (kbps) for speech and audio, and both kilobits per second (kbps) and megabits per second

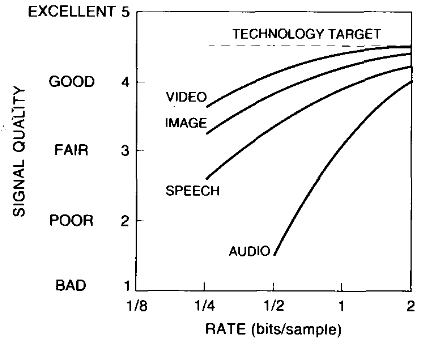


Fig. 5. Current capabilities in the coding of audiovisual information.

(Mbps) in the case of video. All of these numbers can be converted to equivalent numbers in bits per sample based on sampling rates such as the illustrative numbers in Tables 1 and 2. For example, the 4-kbps, 48-kbps, and 15-Mbps rates for 8-kHz–sampled speech, 48-kHz-sampled audio, and 60-MHz-sampled HDTV correspond, respectively, to 0.5, 1.0, and 0.25 bits per sample. In the definition of technology targets in Section II-H, we shall use the normalized unit of bits per sample in the interest of a unified perspective for audio and visual signals.

H. Technology Targets

Figure 5 is a simplified and impressionistic summary of current capabilities in signal coding, expressed in terms of subjective quality as a function of bit rate. The results are derived from a combination of published work [16], [20], [32], [44], [45], [50], [65], unpublished reports, and by collective impressions of experts, especially in the case of image and video signals where formal evaluations of quality are not generally available. Signal quality is measured on a subjective five-point scale ranging from *bad* to *excellent*, as in our earlier description of the *mos* scale.

One of the implications in Fig. 5 is that video signals are the easiest to compress on a bit-per-sample basis. This is attributable to the well-known redundancy in video information in both the spatial and temporal domains of the signal, a property that is also reflected in the extreme low-pass nature of the power spectral density of typical video. By contrast, it is not unusual to encounter relatively flat power spectra in 20-kHz audio. This, combined with the universal expectation of very high levels of quality in entertainment audio, leads to generically lower subjective scores in compressed audio at a given number of bits per sample.

In the category of still images, facsimile documents constitute a special subclass, if we agree to regard text and line graphics, rather than grey-level photographs, as typical fax documents. A half-toned (black–white) document is generally highly compressible. The bit rate for the lossless coding of such a fax document can be typically on the order of 0.1 bit per sample.

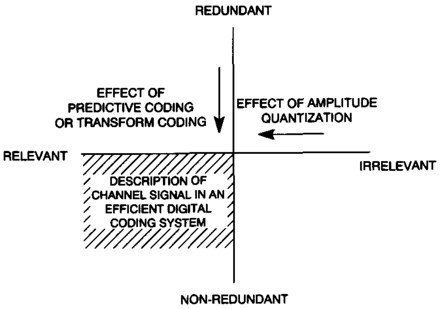


Fig. 6. The Schouten diagram: redundancy and irrelevancy.

As we seek to advance the state of the art as depicted earlier in Fig. 4, it is useful to talk about bit rate targets at which one expects the four signals in Fig. 5 to be digitized with a quality rating such as 4.0 or higher. Without loss of generality or realism, all of these targets can be collectively described by the bit-rate-independent horizontal broken line of 4.5 quality in Fig. 5. Clearly, we are closer to this goal in some signal domains than others. It is also possible that the 4.5 quality goal at rates down to 0.25 b/sample is impossible to achieve in some cases, regardless of coder complexity or processing delay, because of fundamental limits imposed by information theory and the acuity of the human perceptual system. But it is fair to ask the question: as we seek to approach these (sometimes unattainable) levels of high quality at low bit rates, what are the techniques most likely to succeed?

I. Tools of the Trade: Perceptual Coding

There are four fundamental operations that are common to low bit rate signal coding: *prefiltering* to render a signal more "acceptable" to a bit-rate-constrained coder, reduction of signal redundancy in the input signal, removal of irrelevant information in the operation of quantization, and signal enhancement by *postfiltering*. Of these, prefiltering and postprocessing are generally considered to be processes outside of the coding operations *per se*, although the benefits of performing the process can be very significant, as in low bit rate speech coding [19], [20], [44], [123] and low bit image coding [85], [159]. Prefiltering can likewise increase the performance of a compression algorithm. This is accomplished in video coding, for example, by the reduction of camera noise in the input image or by the insertion of an explicit bandlimiting filter. The remainder of this paper will focus on the two operations that are intrinsic to signal coding: *removal of redundancy* and *reduction of irrelevancy* (Fig. 6).

Almost all sampled signals in coding are redundant because Nyquist sampling typically tends to preserve some degree of inter-sample correlation. This is reflected in the form of a nonflat power spectrum. Greater degrees of nonflatness, as resulting from a low-pass function for signal energy versus frequency, or from resonances (in audio) and periodicities (in audio and video), lead to greater gains in redundancy removal. These gains are also referred to as prediction gains or transform coding gains, depending on whether the redundancy is processed in the time domain or in the frequency (or transform) domain.

In a signal compression algorithm, the inputs to the quantizing system are typically sequences of prediction errors or transform coefficients. The idea is to quantize the time components of the prediction error, or the transform coefficients, just finely enough to render the resulting distortion imperceptible, although not mathematically zero. If the available bit rate is not sufficient to realize this kind of perceptual transparency, the intent is to minimize the perceptibility of the distortion by shaping it advantageously in time or frequency, so that as many of its components as possible are masked by the input signal itself. We have used the term *perceptual coding* to signify the matching of the quantizer to the human auditory or visual system, with the goal of either minimizing perceived distortion, or driving it to zero where possible. These goals do *not* correspond to the maximization of signal-to-noise ratio (the minimization of mean squared error).

The parts of a coder that process redundancy and irrelevancy are sometimes separate, as in the above explanation. On the other hand, there are examples where the two functions cannot be easily separated. One example is a vector quantizer that combines inter-sample processing and quantization in a single stage of processing.

Almost all coding systems depend on the complementary interworking of the two basic operations defined above. A notable exception is a pulse-code modulation (PCM) system, based on memoryless coding and quantizing algorithms, where there is no attempt to remove signal redundancy. This simple procedure is adequate for high-quality coding at bit rates in the range of 8 to 16 b/sample, depending on the input signal. On the other hand, low-bit-rate coders, such as those evaluated in Fig. 5, depend heavily on more sophisticated signal analysis, processing delay, and redundancy removal prior to perceptually tuned quantization.

The conceptually orthogonal principles of redundancy reduction and irrelevancy removal were realized fairly early in the field of signal coding (Fig. 6). In this sense, the concept of perceptual coding is not novel. What is significant is that both dimensions in Fig. 6 have advanced considerably over the years, and the interactions between the two dimensions are also getting better understood in coder design.

The next section describes the properties of audiovisual signals and the psychophysics of perception, and thus motivates generic designs of algorithms that perform the dual functions of Fig. 6.

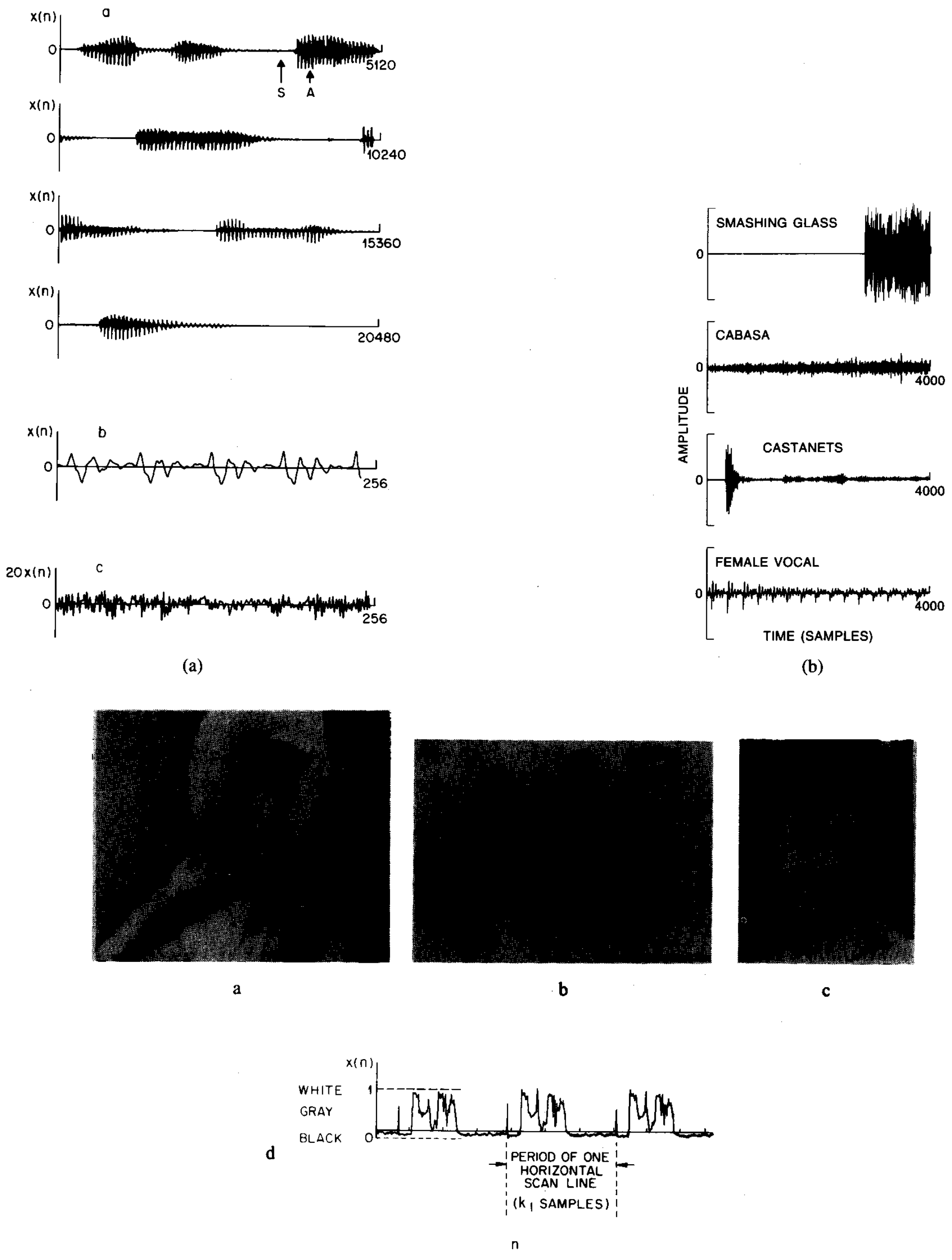
III. Audiovisual Signals, Human Perception, and Time–Frequency Analysis

The input to the digital coder of Fig. 1 is the *source*, the audiovisual signal. After compression (coding followed by decoding), the distorted low-bit-rate version of the signal is evaluated by the *sink* in Fig. 1, the human perceptual mechanism. Common to this mechanism and the coding algorithm are various signal processing and computing functions, prominent among which is time–frequency analysis. The notion of coding algorithms simulating the human perceptual system in their internal optimization loops is a useful one; and in fact, one might measure the sophistication of an algorithm in terms of how well it has evolved in response to the human perceptual model, to the extent human perception is understood. An even more fundamental evolution is associated with the source–sink model itself; at least in the case of speech, it is very likely that the perceptual mechanism (the human auditory system) and the source (the human speech-producing mechanism) have evolved over time to be efficiently reciprocal in some sense. This reciprocity, however, is not exact in any known sense. For example, the frequency responses of the vocal and auditory mechanisms, while being bandpass in each case, are not quantitatively related in any profound way to the best of our current knowledge.

A. Properties of Audiovisual Signals

Audiovisual signals are generally recognized as amplitude plots, amplitude-versus-time waveforms of speech and audio information (Fig. 7(a) and (b)), intensity-versus-space displays of image information (Fig. 7(c)), and intensity-versus-space-and-time displays of video scenes. These *waveforms* reveal significant information about the properties of the signal, and about coder functions that will be needed to utilize these properties for efficiency in compression.

1) Nonstationarity: Common to all of the waveforms of Fig. 7 is the property of *nonstationarity*. In speech, unvoiced signals are characterized by low intensity and short time support; and these signals connect strong, sustained bursts of voiced signals. An unvoiced sound "s" (from the word "salt") appears on row 1 of Fig. 7(a) and on row 6, where it is stretched out in time (to 256 samples or 32 ms), and blown up in amplitude, by a factor of 20. A voiced sound "a" appears on row 1, and part of it is stretched and amplified in row 5. Audio signals likewise have sustained sounds as well as sharp attacks and transients. The female vocal sound in row 4 of Fig. 7(b) is a 4000-sample excerpt of a voiced signal at 48 kHz (83 ms), while the other three audio examples in Fig. 7(b) are noise-like. Image and video signals contain a wealth of segments of flat or slowly changing intensity, as well as edges and textured regions. Parts a, b, and c of Fig. 7(c) are examples of grey-level images and part d is a 3-scan-line excerpt from part c, taken from the portion containing the eyes of *Joanna.* An important function of coding algorithms is to render unvoiced sounds, audio attacks, and edge information faithfully, although these subsignals may represent minority subclasses in respective signals from the point of view of energy or from the viewpoint of the fraction of samples or pixels involved. Adaptive algorithms for quantization, prediction, and bit allocation are designed to improve the performance of coding algorithms with



(a)

(b)

(c)

Fig. 7. Waveforms of (a) speech (b) audio, and (c) image signals. Parts (a) and (c) are after [65].

nonstationary inputs. Perceptual coding provides additional capabilities in terms of handling these signals.

2) Periodicity: There are several sources of periodicity in audiovisual signals. Examples are the pitch periodicities in voiced speech and singing voice (Row 5 of Fig. 7(a), Row 4 of Fig. 7(b)), the periodicities in various forms of instrumental music, and the line-to-line and frame-to-frame periodicities in image and video signals (for example, part d of Fig. 7(c)). In reality, some of the above signals are not exactly periodic, and an important function of a redundancy-removing algorithm is to realize a good prediction performance by keeping track of the changing or evolving periodicities in an adaptive fashion. Examples of such functions are pitch tracking in speech and motion compensation in video. In Row 5 of Fig. 7(a), there are about four pitch periods with a pitch frequency that varies slowly in the neighborhood of 125 Hz (corresponding to about 64 samples at 8 kHz). Accurate pitch estimation

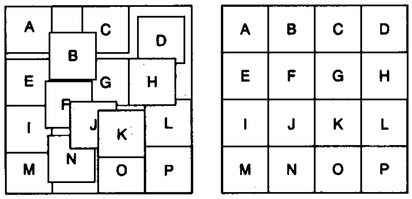


Fig. 8. Motion compensation on a block-by-block basis.

can provide very reliable prediction of signal samples from their counterparts about one period away. In Fig. 8, which depicts two successive image frames, image blocks labeled A, C, E, I, L, M and P get repeated with little change at a period corresponding to one frame period; these blocks are perfectly predicted by their counterparts in a previous frame; the remaining blocks, corresponding to moving parts of the image, are also well predicted by suitably displaced blocks, after motion estimation.

3) Power Spectral Density: In a global or long-time-average sense, audiovisual signals tend to have *lowpass* frequency spectra. When very low frequencies are missing because of a transducer transfer function, as in telephony, the power spectral density will be a *band-pass* function. Short-term frequency analysis, however, reveals parts of audiovisual signals that are either *allpass* or *highpass*; examples are some unvoiced speech sounds and audio and image signals with a predominance of high frequencies, a prototypical example being a checker-board image pattern. On top of the above categorization of the spectral envelope as a function of frequency, it is significant that certain input spectra have powerful forms of local structure. Prominent among these are the *formant* resonances and the *pitch*-related fine structure in the spectra of voiced speech and vocal music (Fig. 9). The upper part of Fig. 9 corresponds to the voiced speech example in Fig. 7(a). In the upper part of Fig. 9, the first two *formants*, or vocal tract resonances occur approximately at 600 and 1200 Hz. The pitch frequency is about 125 Hz as in Fig. 7(a).

Frequency structure has been utilized to great advantage in linear predictive coders (LPC systems) for speech, and subband or transform coders for high-fidelity audio. In the former case, although linear prediction is a time-domain operation, the frequency description has been central to the design and calibration of the redundancy-removing operation. Incidentally, a *white* or flat power spectrum corresponds to an unpredictable time signal, one for which the functional diagram of Fig. 6 collapses to a single dimension, the horizontal axis of quantization.

4) Properties of Color and Stereo Signals: Not included in the above summary are signal-specific attributes such as those of stereo channels in audio and the chrominance channels in color images [10], [12]. We will now comment very briefly on these two issues.

The statistical correlations between the *left–right* audio channels in a stereo pair are easily utilized in a system that is based on mapping the signals to a *sum–difference* pair, which is equivalent to the simplest example of a linear transform coder [65]. It is interesting, however, that this is not by itself a perceptually optimal paradigm, and this is indeed one of the topics of Section VII.

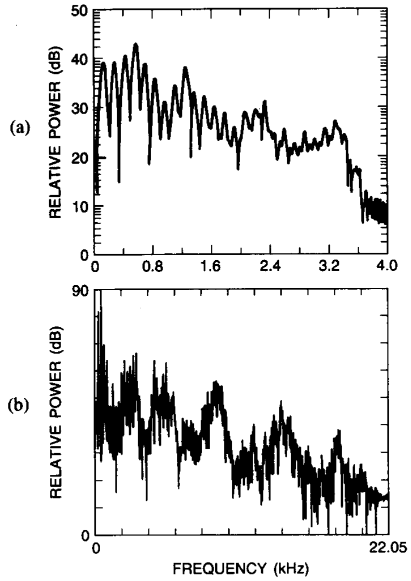


Fig. 9. Power spectral density of (a) voiced speech (after [65]) and (b) vocal music.

The use of so-called YUV and YIQ spaces (rather than the RGB or red–green–blue space) in color image processing is a good example of a paradigm that is efficient both statistically and perceptually. In typical inputs, the UV or IQ (chroma) components have lower energies and lower bandwidths compared to the luminance component Y. This permits the use of subsampling for the chrominance signals as well as the use of coarser quantization compared to luminance. Significant efficiencies in overall bit rate are attained as a result. In a typical color system for low bit rate video, the overall bit rate overhead for chroma can be as low as 10 to 25% of the total bit rate. Perceptual criteria relate to the lower sensitivities to coarse quantization, and to the related observation that additional liberties in chroma quantization are possible in the context of a strong change in the luminance component.

B. Models of the Human Perceptual System

A commonly used model of audition includes a neuromechanical process followed by frequency analysis and detection. A common model of vision incorporates a lowpass filter, a logarithmic nonlinearity, and a multichannel signal-sharpening highpass filter. A biologically correct and complete model of the human perceptual system would incorporate descriptions of several physical phenomena including peripheral as well as higher level effects, feedback from higher to lower levels in perception, interactions between audio and visual channels, as well as elaborate descriptions of time–frequency processing and nonlinear behavior. Some of the above effects are reflected in existing coder algorithms, either by design or by accident. For example, certain forms of adaptive quantization and prediction provide efficient performance in spite of inadequate response time because of temporal *noise masking*, a term that we will describe in some detail in Sections VII and IX. However, even where the matching to the human perceptual mechanism is by design, the matching is intuitive and approximate rather than formal or exact. In other words, at the time of this writing, an *optimal* perceptual coder does not exist. As a matter of fact, a biologically complete model of human perception also does not presently exist. In addition, significant inter-subject variations exist in most examples of human perception.

An example of a perceptual phenomenon that is not directly reflected in compression technology is the possible feedback from higher to lower levels in perception. One example of an effect that is informally reflected in coder design is the nonlinearity model in Weber's law [100]; this is used as a justification for nonuniform quantization and homomorphic coding, as elaborated later. Another example is knowledge about the importance of phase [10], [111]. This is used as a justification for phase-preserving filter banks, especially in spatial image processing; and for avoiding spatial shifts in transitional regions of an image.

Focus in this section is on the time–frequency analyses in human perception, and the desire to incorporate corresponding capabilities and flexibilities in coder modules that try to utilize the phenomenon of *distortion masking or noise masking*. Masking is a complex result of the transducing and neural components of perception. It is highly adaptive and refers to the perceptibility of one signal in the presence of another in its time or frequency vicinity. In coding, one of the signals is the input, the second is the distortion in the low bit rate coding of it, or in the communication of it over a noisy channel or a lossy network.

The basic time–frequency analyzers in the human perceptual chain are described by *bandpass filters* both in audition and vision. Table 3 [132] is a model consisting of 25 cochlear filters, with a bandpass width that increases monotonically with increasing frequency. The 25 subbands of Table 3 are also referred to as *critical bands* [38], and their significance is explained in Section VII. The audio frequency response in Fig. 10 is a function of frequency (in Hertz) and loudness level. Human auditory acuity tends to decrease at low and high frequencies, particularly when the loudness is low. This is a composite result of the cochlear filter bank, the frequency response of the ear canal, and other (partially understood) phenomena in human hearing. The visual response in Fig. 11 is a function of spatial frequency (in cycles per degree, cpd) as well as temporal frequency (Hertz). In each case, the response shows distinct peaks in individual bandpass characteristics.

Bandpass filters in audiovisual perception are sometimes reflected in coder design and telecommunication practice in the forms of "rules of thumb." For example, if the viewing distance in high-definition television is 8 ft and the image resolution is 1200 × 800 pixels, it is necessary that the physical image (screen) size is 5 ft in diagonal measurement so that important high frequencies in the image correspond to numbers lower than a low-sensitivity point such as 25 cpd in the spatial response curve of Fig. 11. In Sections VII and VIII of this paper, we shall attempt to incorporate bandpass filter information more explicitly; we have been successful in doing this in the calculation of the distortion-masking models in audio and still-image coding.

Table 3 Crytical Band Centers and Edge Frequencies from Scharf (1970)

Band Number (Hz)	Lower Edge (Hz)	Center (Hz)	Upper Edge (Hz)
1	0	50	100
2	100	150	200
3	200	250	300
4	300	350	400
5	400	450	510
6	510	570	630
7	630	700	770
8	770	840	920
9	920	1000	1080
10	1080	1170	1270
11	1270	1370	1480
12	1480	1600	1720
13	1720	1850	2000
14	2000	2150	2320
15	2320	2500	2700
16	2700	2900	3150
17	3150	3400	3700
18	3700	4000	4400
19	4400	4800	5300
20	5300	5800	6400
21	6400	7000	7700
22	7700	8500	9500
23	9500	10500	12000
24	12000	13500	15500
25	15500	19500	

A particularly interesting aspect of the signal processing model of the human system is nonuniform frequency processing. The so-called critical bands in hearing and vision are nonuniform. To incorporate this in coder design, it is necessary to use masking models with a nonuniform frequency support (Section VII). It is also necessary to recognize that high-frequency signals in audiovisual information tend to have a short time or space support, while low-frequency signals tend to last longer. An efficient perceptual coder therefore needs not only to exploit properties

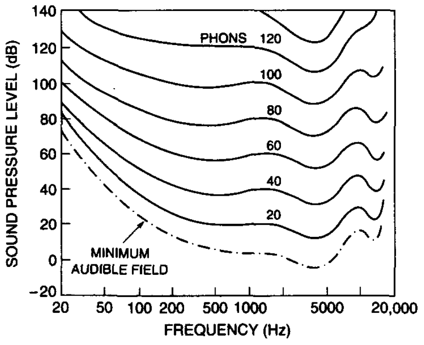


Fig. 10. Frequency response characteristics of the human auditory system as a function of loudness.

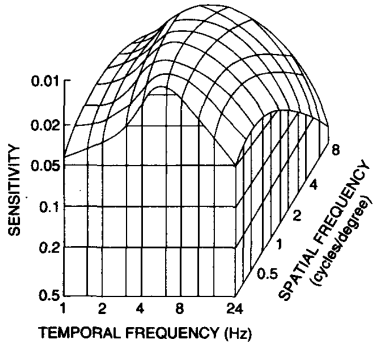


Fig. 11. Spatio-temporal sensitivity of the human visual system (after [100]).

of distortion masking in time and frequency, but it also needs to have a time–frequency analysis front-end that is sufficiently flexible to incorporate the complex phenomena of distortion masking by nonstationary input signals. All of this is in contrast to the classical redundancy-removing coder driven purely by considerations of *minimum mean squared error* (mmse), mmse bit allocation, or mmse noise shaping matched to the input spectrum (rather than to the composite of input spectrum and perceptual spectrum).

IV. Filterbanks and Transforms in Signal Processing

Predictive methods as well as subband and transform techniques have been used as bases for frequency-dependent signal coding. Subband filterbanks and transforms provide direct control of frequency or frequency-related properties in the input signal.

A. The Quadrature Mirror Filterbank and the Discrete Cosine Transform

Figure 12 depicts examples of subband decompositions in one, two, and three dimensions. The 1D analysis is used for speech and audio analysis, and sometimes in image processing. The 2D and 3D decompositions are particularly appropriate for image and video analyses, respectively. It is assumed that the filterbanks used for these decompositions satisfy requirements on reconstruction error, processing delay, and interband interference. In typical examples of 2D and 3D filterbanks, the structures are realized by separable, identical operations in the multiple dimensions. Quadrature mirror filterbanks (QMF) and related structures are very efficient designs in that the filters in the bank are of reasonably low order to permit practical implementation. Typically, finite impulse response (FIR) structures with 16 to 64 taps [66] are used. Passband ripple is nearly zero with a 64-tap filter. The interband aliasing due to overlapping filter responses is controlled in a way that permits subsequent cancellation of it in the synthesis filter. In the absence of quantization, the analysis and synthesis filterbanks constitute a unity or near-unity operation [149]. Efficient signal compression results when subband signals are quantized with subband-specific bit allocation and quantization, based on input power spectrum and the model of perception.

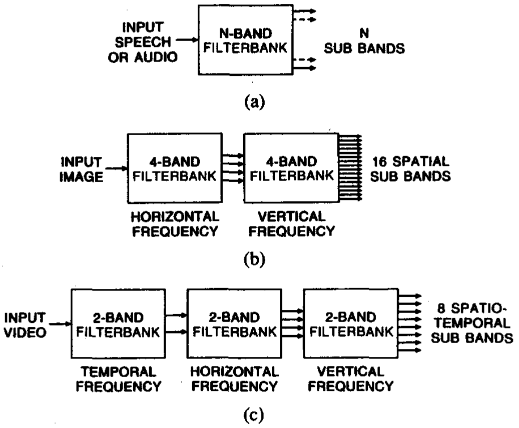


Fig. 12. Subband decompositions in one, two, and three dimensions of frequency.

The Discrete Cosine Transform (DCT) is well known for its efficiency in speech and image compression [1], [65]. The image blocks in Fig. 13 are the 64 basis images of an 8 x 8 2D transform. They capture a range of spatial (horizontal and vertical) frequencies from the lowest to the highest, with 62 intermediate combinations of horizontal and vertical frequency. The 2D Discrete Cosine Transform (DCT) decomposes the image into components that are analogous to the vertical and horizontal frequency components in the 2D filterbank of Fig. 12(b). In the absence of quantization, the 2D DCT and the identical operation of the inverse DCT at the receiver (2D IDCT) constitute a unity operation. Efficient signal compression results when the transform coefficients are quantized with coefficient-dependent bit allocation and quantization.

In subband and transform coders, the number of frequency components in the decomposition (subbands or coefficients) is a compromise between prediction gain (resulting from variable bit rate coding of the unequal components in the nonflat spectrum of the redundant signal) and practical considerations such as complexity and processing delay.

Subband and transform coders provide a natural framework for variable rate coding, embedded coding, unequal bit allocation, and unequal channel error protection. The

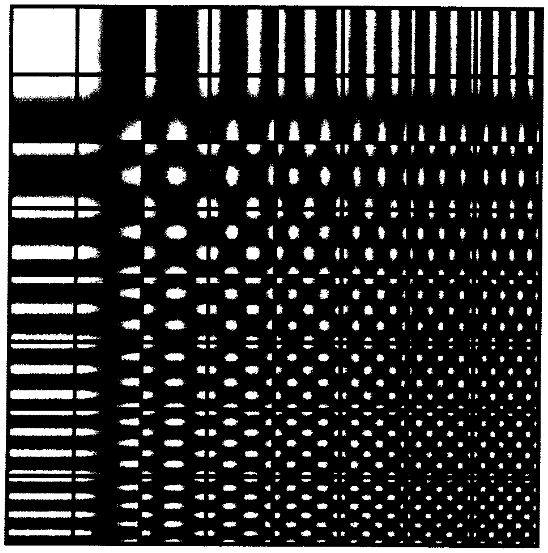


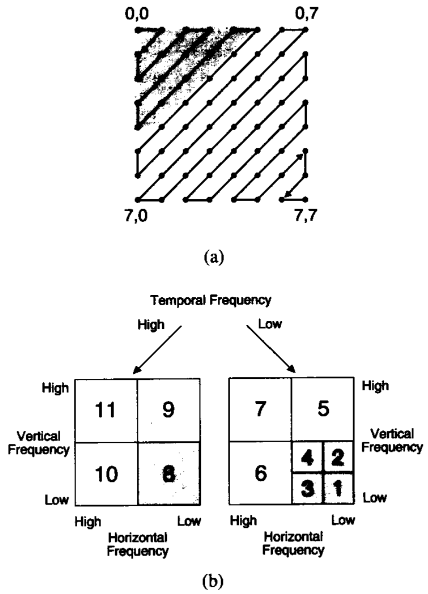
Fig. 13. Basis vectors of a 2D 8 × 8 Discrete Cosine Transform (DCT).

CCITT 64/56/48 kbps standard for wideband speech [16], [96] is a variable rate coder based on a 2-band coder with subbands (0-4 kHz) and (4-8 kHz). The higher subband is sometimes quantized coarsely to accommodate 8 or 16 kbps of data in the 64-kbps channel. The ISO standard for 24-kHz audio is a three-layer standard [98], [108]. The first two layers are based on a 32-band decomposition of the 24-kHz signal, with each subband occupying 750 Hz. Nonuniform subband decompositions are sometimes used for efficient perceptually tuned coding. One example is the 3D subband decomposition of Fig. 14(b). In an embedded coder, the shaded subbands constitute an example partition for *essential* information while the unshaded subbands could be regarded as *enhancement* information. In a 2-layer embedded coding system as in an ATM packet network [77], [97], [109], [151], the enhancement layers are much more likely to be discarded in the event of network congestion. The above hierarchical features also apply to transform coding. The shaded area in Fig. 14(a) represents one example of a partition for *essential* 2D DCT coefficients.

Frequency-domain coders (subband, transform and related multiresolution systems) also offer an excellent framework for *progressive transmission*. In an application like telebrowsing, it may be useful to obtain a rough version of an image first, using minimal communication resources; and to request a finer version only if an image needs to be scrutinized further. In this application, progressive transmission of frequency content is subjectively more useful than the *sequential* top-to-bottom transmission of pixel rows (Fig. 15).

B. The Mathematics of QMF and DCT Analyses

The coefficients describing the QMF analysis filter are reused with a trivial modification in the QMF synthesis filter. The DCT and IDCT equations are likewise very similar. The operations describing the QMF filter and the DCT are described by the following equations:



(a)

(b)

Fig. 14. Embedded coding in (a) transform and (b) subband frameworks.

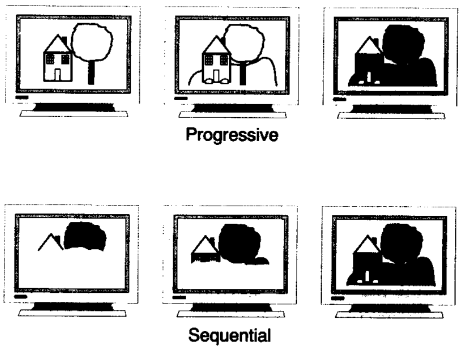


Fig. 15. Progressive versus sequential transmission.

QMF:

$$h_\ell(n) = h_u(n) = 0, \qquad \text{for} \quad 0 > n \geq N$$

$$h_\ell(n) = h_\ell(N-1-n), \qquad n = 0, 1 \cdots N/2 - 1$$
$$h_u(n) = -h_u(N-1-n), \qquad n = 0, 1 \cdots N/2 - 1$$

$$h_u(n) = (-1)^n h_\ell(n))$$
$$\left|H_\ell(e^{j\omega})\right|^2 + \left|H_u(e^{j\omega})\right|^2 = 1$$

DCT:

$$F(u) = \sqrt{\frac{2}{N}}\alpha(u) \sum_{n=0}^{N-1} x(n) \cos\frac{(2n+1)u\pi}{2N}, \qquad u = 0, 1 \cdots N-1$$

$$\alpha(0) = 1/\sqrt{2}; \quad \alpha(u) = 1 \quad u \neq 0$$

IDCT:

$$x(n) = \sqrt{\frac{2}{N}} \sum_{k=0}^{N-1} \alpha(u) F(u) \cos \frac{(2n+1)u\pi}{2N},$$
$$n = 0, 1 \cdots N-1$$

2D DCT:

$$F(u,v) = \frac{2}{N} \alpha(u)\alpha(v) \left[\sum_{m=0}^{N-1} \sum_{n=0}^{N-1} x(m,n) \cdot \cos \frac{(2m+1)u\pi}{2N} \cos \frac{(2n+1)v\pi}{2N} \right]$$

$$\alpha(u) = \alpha(v) = 1/\sqrt{2}; \alpha(u) = \alpha(v) = 1 \text{ if } u \neq 0, v \neq 0$$

2D IDCT:

$$X(m,n) = \frac{2}{N} \left[\sum_{u=0}^{N-1} \sum_{v=0}^{N-1} \alpha(u)\alpha(v) F(u,v) \cdot \cos \frac{(2m+1)u\pi}{2N} \cos \frac{(2n+1)u\pi}{2N} \right].$$

C. The Duality of QMF and MDCT

The basic 2-band QMF system (Fig. 16(a)) is based on the division of a frequency band into two contiguous but overlapping subbands. The broken line shows the characteristic of a high-pass filter that is the mirror image of the solid-line low-pass filter characteristic. These filters intersect at the −3-dB point of either characteristic. The extent of overlap is a decreasing function of the number of filter taps. But the allowing of a nonzero overlap simplifies filter design. Frequency aliasing is caused in QMF analysis by sampling each of the two bands (lower and upper bands) in the QMF split at twice the *nominal* bandwidth (rather than twice the actual, say −90-dB, bandwidth). However, with a special design of QMF filters, the process of QMF synthesis provides cancellation of this aliasing if the quantization noise inserted in the system is zero [25], [34], [65], [66], [149].

Transform coding as used in early examples from telephone speech coding uses a Discrete Fourier Transform or DFT, calculated via an FFT, or a Discrete Cosine Transform. The equivalent frequency response of a transform channel depends on the window used. Commonly used window functions are the rectangular and sine-taper functions. With a rectangular window as used in [1], the analysis/synthesis system is critically sampled, but the system suffers from poor frequency resolution and blocking noise, which can be regarded as poor filtering and antialiasing of the noise introduced into the transformed coefficients. Overlapped windows allow for better frequency response functions but carry the penalty of additional values in the frequency domain which have to be coded given a certain number of time domain samples. The system is not critically sampled. For use in coding, a compromise between large overlap giving better frequency resolution and low overhead due to overlap has to be found. One widely used compromise is an overlap of 1/16 of the window length. The modified discrete cosine transform (MDCT) is a clear exception to this practice.

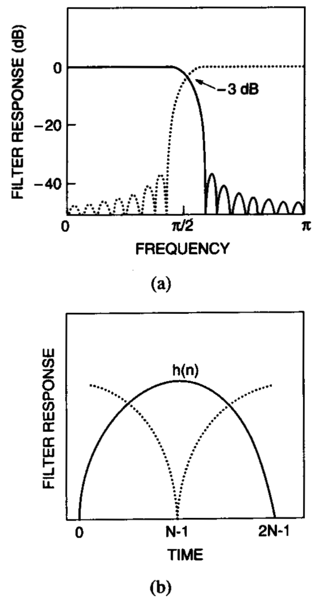


Fig. 16. (a) Quadrature mirror filter and (b) overlapped transform in MDCT.

The modified DCT system [68], [91], [98], [119] is a dual of the QMF approach in that it permits an overlap between successive transform blocks in the time domain, but decimates the resulting sequence to maintain the original sampling rate. The overlapping in time is described by the window function $h(n)$ of Fig. 16(b) and the (dotted) time-displaced versions of it. In place of the frequency aliasing in the QMF system, the MDCT exhibits time-domain aliasing because of the decimation. But this aliasing is canceled by the inverse MDCT process in the receiver due to the design of the DCT basis vectors and the analysis window. The overlap in the MDCT is 50% and the decimation rate is 2:1.

The equations describing MDCT are given by

MDCT:

$$F(u) = \sum_{n=0}^{2N-1} h(n) x(n) \cos \left[\frac{\pi}{2N} (2u+1)(2n+1+N) \right]$$
$$u = 0, 1 \cdots 2N-1$$

$$h^2(N-1-n) + h^2(n) = 2$$
$$h^2(N+n) + h^2(2N-1-n)^2 = 2$$

$$0 \leq n < N.$$

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

The MDCT is the dual of a 2-tap Haar-QMF filter, obtained by using $N = 2$ in the QMF equations mentioned earlier.

QMF and MDCT systems are qualitatively dual in the sense of time versus frequency aliasing (and aliasing cancellation); they are exactly dual in the case of the 50% overlap MDCT and a 2-tap QMF filter (which has a 50% frequency overlap). However, the QMF and MDCT systems exhibit different properties in the context of a specific overall coder. For example, operations such as signal anticipation and temporal bit allocation can be used to some extent to control quantizing distortion and the consequent phenomenon of uncancelled time-domain aliasing in the MDCT system. This in turn permits the use of the 50% time overlap which provides a smooth handling of the time process as well as a simple decimator design. In QMF design on the other hand, one still prefers, in the current state of the art, to employ a small spectral overlap, within the constraints on filter complexity. This in turn, implies a relatively discontinuous handling of the frequency process.

Finally, in both MDCT and QMF, the time–frequency characteristic in signal analysis is a compromise between a match to sustained, stationary inputs and the ability to handle input discontinuities in time or frequency. In the coding of stationary segments, the approach is to use a high-resolution input decomposition to maximize coding gain, even if the decomposition does not mimic the human filter. This is particularly true in audio coding. In the coding of nonstationary segments, such as a sharp attack in an audio signal, it is more useful for the filterbank to mimic the auditory mechanism.

D, Criteria for Filterbank Design

The sampled filterbank used to decompose the time-domain input data into components with the desired spectral and time-domain resolution is crucial for the overall performance of a perceptual coding system. Several competing design goals have to be met:

- The spectral decomposition and subsampling process should be invertible, i.e., the filterbank should be of the *perfect reconstruction* type. It is also desirable that the filterbank and its inverting process both maintain a high degree of frequency selectivity, in order to make the application of the perceptual threshold simple.
- The analysis system, consisting of the analysis filterbank and the subsampling, should be critically sampled or nearly critically sampled. The number of spectral components per time unit determines the number of values to be quantized and coded and therefore should be as low as possible. The optimum in this respect is a critically sampled analysis/synthesis system.
- The bandwidth of the filterbank should be equal to or narrower than the width of a narrowest critical band. This makes it easier to control the perceptibility of the quantization noise. Simultaneously, the time window of the filterbank must be controlled in order not to introduce noise components over a time window wide enough that masking constraints are violated. In general, most uniformly spaced filterbanks cannot meet both of these constraints due to the widely varying width of critical bands with frequency.
- Very high frequency resolution is desirable to take advantage of the *transform gain*, which is the frequency-domain equivalent of the prediction gain. The transform gain is highest for signals with a low spectral flatness. These signals need to be coded very accurately because there is very little masking of the quantization noise by such inputs.
- The time-domain resolution of the filterbank should be low enough to mask the spreading in time of the error signal. In a subsampled system, the lower bound for the time resolution is given by the Nyquist bound. The quantization noise in the frequency domain is spread in time or space according to the actually used filterbank. This effect causes *pre-echo* in audio processing (Section VII).

Considerations of signal nonstationarity and perceptual distortion criteria have resulted in increasingly sophisticated demands on signal analysis. In particular, techniques that provide flexible combinations of time support and bandwidth represent a powerful generic tool for efficient coding. The discrete Fourier transform and a uniform-bandwidth quadrature mirror filterbank are well-understood and widely used analysis tools. But in their simplest forms, they lack the flexibility for time–frequency analysis mentioned above. QMF trees with unequal-bandwidth branches, as well as subband–DFT hybrids, are relatively newer structures with more flexible features. So are *wavelet* filters [31], [125], [153].

1) Wavelets: Unlike the basis vectors of a DFT (sinusoids and cosinusoids of various frequencies and constant time support), the wavelet filter structure is characterized by a shorter time support at higher frequencies, and a longer time support at lower frequencies, a direct result of a dilating operation that is a basic component of wavelet design [153]. The time–frequency characteristic of a wavelet filterbank is a natural match to some of the properties of audio-visual information: high-frequency events often occur for a short time and stand to benefit from a finer resolution in time analysis, while low-frequency events are often sustained in time, and require less frequent sampling in time. The wavelet approach, especially if used in a time-varying framework, may therefore offer powerful forms of adaptive analysis in a more basic sense than a nonuniform frequency-band QMF system or a variable window-length MDCT system. Wavelet transforms provide the additional feature of perfect reconstruction, a property not generally offered in conventional methods of analysis. Conventional methods, however, do offer the property of *almost-perfect* reconstruction which is adequate for many low bit rate applications.

Wavelet filtering is a promising analytical tool and applications of it are also beginning to emerge. What is still lacking, however, is a thorough understanding of what wavelets can do for coding that the more sophisticated examples of conventional analyses cannot; and as we seek to apply wavelet (or nonwavelet) tools to low-bit-rate coding, attention must necessarily shift to the yet-untouched problems of uncancelled aliasing, and the computationally intensive but extremely important notion of a signal-adaptive filterbank.

E. *Effect of Quantization Error*

The assumption of zero aliasing is a conspicuous simplification in almost all of subband coding literature. Techniques such as quadrature-mirror filtering and the modified discrete cosine transform can provide perfect cancellation of aliasing in the absence of quantization errors. This ideal solution never occurs in a practical coding situation, and the assumption of the ideal case becomes increasingly inappropriate at lower bit rates because of a corresponding increase of quantization error. Recent work has given us a fairly good understanding of filterbanks that provide zero or near-zero reconstruction error in the absence of quantization. However, the design of a filterbank that minimizes the combined effect of quantizing and aliasing errors is an unsolved problem. Here again, a rigorous optimization is extremely intractable in our current state of knowledge, but we sorely need at least partial solutions. A preliminary approach to the problem is described in [82].

V. Quantization

The bit rate in the digital representation of an analog signal is determined as the bit rate of the (primary) quantizing system in a coder, together with any extra information that the coder may need to signify control signals such as LPC parameters in speech coding and motion vectors in video coding. Such side information is also digitized by means of a (secondary) quantizing system.

The (primary and secondary) quantizing systems consist of either a scalar or a vector quantizer. This is followed (especially in the case of scalar quantization) by an entropy coder such as a Huffman or Ziv–Lempel coder that exploits residual redundancies *after* quantization. Since the entropy coder is a *noiseless* or *information-lossless* operation in a mathematical sense, the quantizer is the part of the overall coder where perceptual coding is explicitly realized.

In this section, we trace the history of quantization in a nonexhaustive manner, and point out the increasing incorporation of perceptual cues as coding systems have become more sophisticated and efficient.

A scalar quantizer is a memoryless device by definition. However, the statistical and perceptual effects of quantization are not adequately described by instantaneous properties. For example, the power spectral density of quantization noise, or equivalently, its autocorrelation function, is a function of the quantizer as well as the input signal. The perceptual impact of quantization noise depends on its spectral distribution and, as we shall see later in discussions of masking, it depends on its relationship to the spectral distribution of the input signal as well. The ratio of the areas under the two spectral densities, the overall *signal-to-noise ratio* (SNR) is a partially useful descriptor of performance for significantly overcoded systems. But with critically coded low-bit-rate signals, the SNR is not only inadequate but can also be misleading in terms of perceptual significance.

1) The PCM Quantizer: Analog signals are stored on the computer as R-bit PCM codes. For example, R may be 16 b/sample for audio and 8 b/sample for the luminance and chrominance components of color images. If these PCM codes are considered to be high-quality representations, it means that the additional resolution provided by an $(R + \Delta R)$-bit PCM system $(\Delta R > 0)$ is perceptually irrelevant (while if $\Delta R < 0$, the signal representation may be considered inadequate). Thus if the coder is constrained to be no more complex than a memoryless quantizer, the R-bit quantizer provides perceptually critical coding of the input at that bit rate.

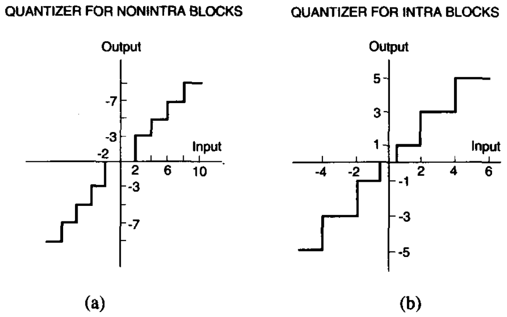


Fig. 17. Midtread quantizers for DCT coefficients in (a) interframe and (b) intraframe coding (ISO-MPEG standard).

2) Nonuniform and Midtread Quantizers: Although the theory of *mmse* quantization is extensive, quantizer designs are often guided by *non-mmse* criteria that make more intuitive sense from a perceptual viewpoint. One example is a nonuniform quantization algorithm for a differential PCM image coding system where the quantizer characteristics are designed to maximize the masking of distortion by luminance changes (the input to the differential coder). A better known and more widely practiced example is that of a midtread quantizer with an odd number $(N - 1)$ of output levels, including zero, sometimes with an extra-wide dead zone near zero. These features, while providing a nonminimum mean squared error, avoid the perceptually unpleasant oscillations between the smallest positive and the smallest negative output levels. Examples of such quantizers are the mid-tread A-law PCM code for high-quality speech [65] and the mid-tread quantizer with dead zone for quantizing DCT coefficients in interframe video coding [85], [100], [156] (Fig.17). The *intra* and *nonintra* categories in Fig. 17 refer to image frames that are coded with *intraframe* (spatial) coding or *interframe* (temporal) coding (Section IX).

3) The Preference of Overload to Granularity in Quantization: Another example of a non-mmse algorithm is one where quantizer design (quantizer step size, characteristic, and/or adaptation algorithm) favors a greater predominance of overload distortion compared to granular noise. This is well understood in deltamodulator designs [65], [141] which favor significant amounts of *slope-overload* distortion (Fig. 18). Such designs are suboptimal from an mse viewpoint but provide the best balance of distortion types from a perceptual viewpoint. To use the formal language of later parts of this section, the reason why the human

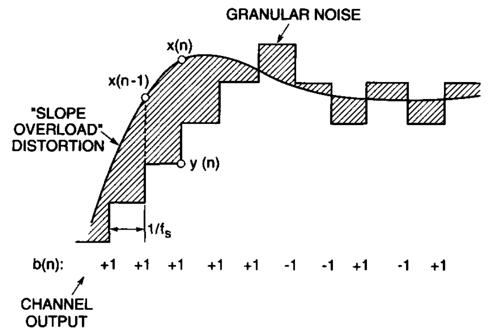

GRANULAR NOISE
x(n)
x(n-1)
"SLOPE OVERLOAD" DISTORTION
y (n)
1/f$_s$
b(n): +1 +1 +1 +1 +1 -1 -1 +1 -1 +1
CHANNEL OUTPUT

Fig. 18. Slope overload distortion and granular noise in deltamodulation (after [65]).

perceptual system is relatively sensitive to granular noise is that the near-zero-slope input (in the right half of Fig. 18) is a poor masker of the high-frequency noise caused by the succession of positive and negative steps in the deltamodulator output.

4) The Use of Dither in Low-Bit-Rate Quantization: In low-bit-rate PCM coders, the quantizing distortion tends to be highly structured, rather than random. This is reflected by significant input–distortion correlations (as in the speech coding example of Fig. 19(a), Row 3), and by *contouring* distortion (as in the right image coding example of Fig. 19(b)). Dithering is a technique where random noise (typically with a small amplitude support) is added to the input *prior to quantization.* The random noise can be subsequently removed from the quantizer output in an ideally synchronized encoder–decoder system; or, as is usual in image coding, there may be no such correction at all. In the latter case, the SNR of the overall system is unchanged. In the former, the SNR actually decreases because of dithering. But in both cases, dithering helps to break up signal-dependent patterns in the distortion and makes the coarse quantizer more palatable. This is seen in the reduction of input-distortion correlations in speech quantization; the resulting perceptual gain is confirmed in subjective tests. In Fig. 19(a), $x(n)$ and $r(n)$ refer to waveforms of speech and quantization error, and R refers to the number of PCM bits. Row 3 of the figure corresponds to quantization without dithering, while Row 2 refers to quantization with dithering. The reduction of input-distortion correlation is clear in Row 2, especially for $R = 1$ and $R = 5$. In the image example (Fig. 19(b)), the bit rate is 3 b/sample. The *right* picture has no dithering. In the *left* image, the inclusion of dithering increases the mean square value of the distortion, but the breaking up of the contours in the image provides a better chance for the input image to mask the distortion, or at least decrease its visibility.

Dithering is an integral part of analog-digital converters (especially for high fidelity audio) and image halftoning. It is a precursor of more advanced systems for noise-shaping.

5) Weighted Distortion: The SNR criterion of quantizer performance can be made perceptually more meaningful by weighting it in some way (Fig. 20). For example, distortion components can be weighted by local input amplitude to reflect diminished sensitivities to errors in low-amplitude input segments. Such weighting can be used to show the desirability of the midtread quantizer of Fig. 17. It is also the principle behind the segmental SNR measure used in speech coding [65].

Perhaps the most important weighting of distortion in the generic system of Fig. 20 is a suitable form of time–frequency weighting. This is not surprising, given the discussion in Sections III and IV. Frequency-weighted distortion has been particularly valuable, for transparent as well as nontransparent coding.

6) Vector Quantization: One example of error weighting is an image vector quantizer which places heavier emphasis on interblock differences, and a relatively lower emphasis on intrablock distortions in a vector quantizer with memory. Such error weighting leads to code-vector selections that compromise the intrablock SNR, but create a perceptually more pleasing low-bit-rate image characterized by smooth interblock transitions [73]. In the example of Fig. 21(a), the tendency is to select codevectors that maximize smooth vertical transitions from a vertical neighbor and smooth horizontal transitions from a horizontal neighbor, with both neighbors being previously quantized with a vector quantizer using the same codebook that is available to the current input.

The higher frequency subbands in 3D coding are dominated by edge-like information of low energy and high perceptual value for scene intelligibility and motion rendition. An efficient method for reproducing the edge-like information at very low average bit rates is the technique of *geometric vector quantization* [117]. In the simplified example of Fig. 21(b), the codebook consists of codevectors that can reproduce horizontal, vertical, and diagonal edges, as well as a null vector for the representation of frequently occurring areas of constant or near-constant grey level. The concept is very close to that of *visual pattern image coding* [18].

7) The Homomorphic Model for Perceptual Coding: In a sense, this paradigm is an extension of the distortion-weighting principle. However, rather than weighting the distortion, the system weights the input and transforms it into a (*perceptually flat*) domain where an unweighted error is useful. As a result, the quantizer itself can be very simple in structure and in terms of cues for optimization. Simple examples of input weighting are the logarithmic compressor for speech quantization [65], and the similar logarithmic nonlinearity following Weber's Law. A generic homomorphic, psychovisual coder [51], [55], [143] using nonlinear preprocessing is shown in Fig. 22. The nonlinearity is part of a model for the Human Visual System (HVS), and the coder includes a gamma correction factor to allow for camera nonlinearity. What is significant in the system is that the input and distortion (introduced in the quantizer Q) go through different weightings before reaching the eye; and emphasis is on optimizing the HVS model rather than the quantizer Q. Perhaps equally significant is the fact that the paradigm of Fig. 22 is a one-signal model, rather than

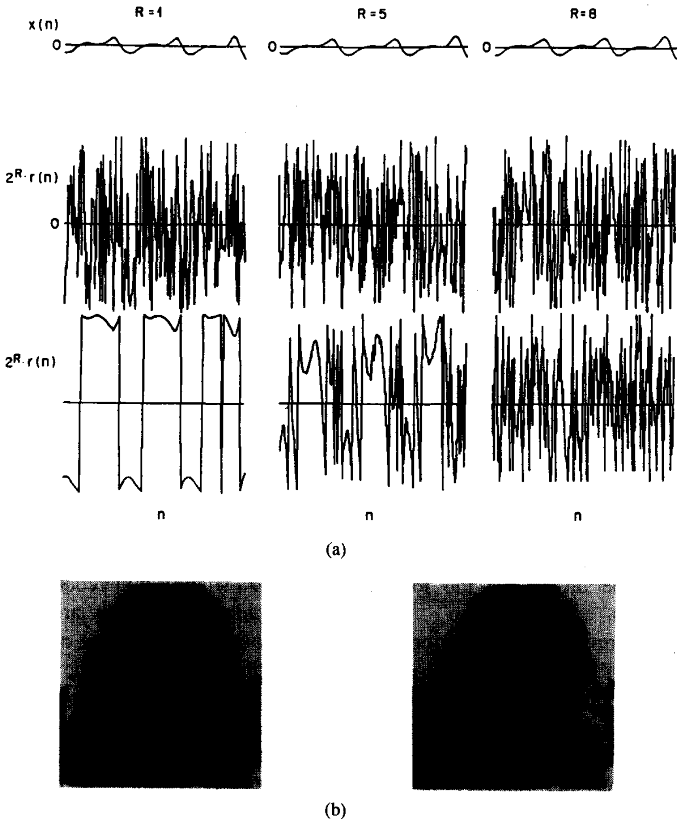


Fig. 19. The destructuring of distortion by the use of dithering in the quantization of (a) speech [65] and (b) image signals (after [126]).

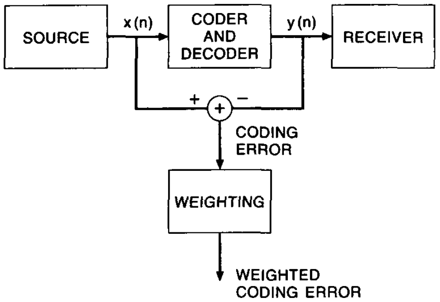


Fig. 20. Weighted distortion in signal coding.

a two-signal masking model; as such, it does not provide a specific mechanism for driving local distortion components to zero. Rather, it tends to minimize noise visibility by the use of a global perceptual model. This global model is also a static one, although dynamic extensions have been proposed recently. Finally, the homomorphic model is not necessarily tuned to the characteristics of a specific compression algorithm.

We next propose a model that is dynamic, local, coder-tuned, and naturally suited to the explicit use of what we know about distortion masking in audition and vision.

8) A Paradigm for Perceptually Lossless Low-Bit-Rate Coding: Figure 23 defines a perceptual coding methodology that has recently produced very promising results in audio

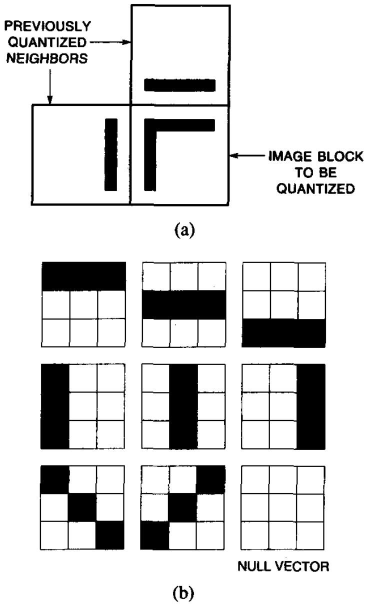


Fig. 21. (a) Adaptive vector quantization with an interblock distortion metric and (b) geometric vector quantization.

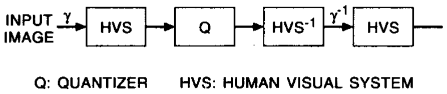


Fig. 22. The homomorphic model for incorporating the human visual system (HVS) in coding (after [51], [143]).

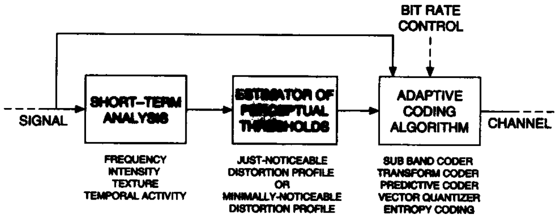


Fig. 23. A paradigm for adaptive perceptual coding.

and image coding. The methodology provides a framework for perceptually lossless coding at the lowest possible bit rate within the constraints of a given coding algorithm. It also provides a framework for perceptually optimum performance when the available bit rate is too low to provide transparent compression.

The first box in Fig. 23 performs a short-time or spatiotemporally local analysis of the input audio or visual signal and extracts several properties such as frequency, intensity, texture, and temporal activity. These local properties are then used in the second stage of the process to derive a perceptual distortion threshold. This threshold can be a function of time, space, or frequency. It expresses a critical distortion profile in the sense that if the distortion caused by the compression algorithm is at or below the threshold at all points in time, space, or frequency, the degradation in signal quality is imperceptible. The critical distortion profile will be called the *just noticeable distortion* (JND). A supra-threshold generalization of it, suitable for nontransparent (but still perceptually optimum) coding will be called the *minimally noticeable distortion* (MND).

The mapping from local properties to the JND profile is done in real time in general, and by necessity in a compression algorithm for two-way communication. The knowledge needed for the mapping is derived on the other hand from off-line experimentation with a large number of subjects performing a coder-specific perceptual task (as in the perceptual image coding algorithm of Section VIII). Alternatively, the knowledge for the dynamic JND derivation may come from adapting results from psychophysical literature to the particular coding algorithm in empirical procedures (as in the perceptual audio coding algorithm of Section VII). In either method, inter-subject variations are addressed by defining a sensitivity to represent $P\%$ of the test population. Values of P may be 50, 75, or 95, for example, depending on how conservative the algorithm is desired to be. In the perceptual experiments described in this paper, we use the 95% criterion. Other sources of variation come into play in describing perceptual effects, such as the effect of viewing distance in image coding. The method of Fig. 23 is, therefore, a generic methodology rather than a specific design.

MND profiles are also derived from off-line information (as in the speech coding examples of Section VI). An alternative, implicit way of deriving them, especially at bit rates that are high enough to permit near-transparency, is to upshift the JND profile until the desired bit rate is achieved, at the cost of undercoding some parts of the input signal.

Given the JND or MND information, the rest of the coding algorithm is in principle straightforward, and it is the function of the third box of Fig. 23. A typical operation at this point is *adaptive bit allocation* that is steered by the JND or MND cues. The short-term analysis implied in these functions precedes the quantization function in the third box of Fig. 23, and this analysis can be either the same as that in the first box of the figure, or, in general, different.

The overall result of the paradigm in Fig. 23 will be a *variable bit rate*, *constant-quality* algorithm. If needed, feedback from a bit buffer can be used to perturb the algorithm into a *constant bit rate, variable-quality* method. Hopefully, in this variation of the coder, most of the signal will be slightly overcoded and only some of the signal will be slightly undercoded. The desired constant bit rate is realized typically by an iterative process which tends to increase the complexity of the perceptual coder.

In the critically coded, variable bit rate mode (which is the only known *optimal* mode of the system of Fig. 23), the total number of bits needed to encode the signal is really a fundamental limit: the lowest rate at which transparent coding is possible. We call this the *perceptual entropy* in deference to the information-theoretic terminology of entropy [9], [137], [138]. Two qualifications are needed in this context. First, these entropy estimates are only as good as the subjective data used to estimate JND profiles; hopefully, as our knowledge about noise-masking improves, lower entropies can result. Second, the perceptual entropy is input-specific and it varies from segment to segment in a nonstationary signal. It is useful in these cases to talk about the long-time-averaged perceptual entropy, as well as histograms of segmental perceptual entropy. Figure 24 is a sample collection of such histograms. The two modes in the telephone speech histogram correspond to voiced and unvoiced speech (left and right modes). The nonzero probability of zero entropy refers to silent blocks in speech. Similar spikes of zero entropy are seen for pop music and orchestra. The entropies illustrated for speech and audio are idealized numbers that serve as lower bounds on actual bit rates. The bit rate information in the image example is based on real quantizers with real Huffman coders following them.

In constant-bit rate coding systems designed for transparent reproduction of the signal, the encoder needs to operate at the peak of these histograms rather than at the average of perceptual entropy. The long tails of the histograms are particularly significant for the speech and audio examples. It will be useful to compare the results of Fig. 24 with the bit rates discussed in Figs. 5 and 4, remembering that no explicit claims about transparency are made in those pictures (except at the value of *mos* equal to or greater than about 4.5, at which point the bit rates needed according to

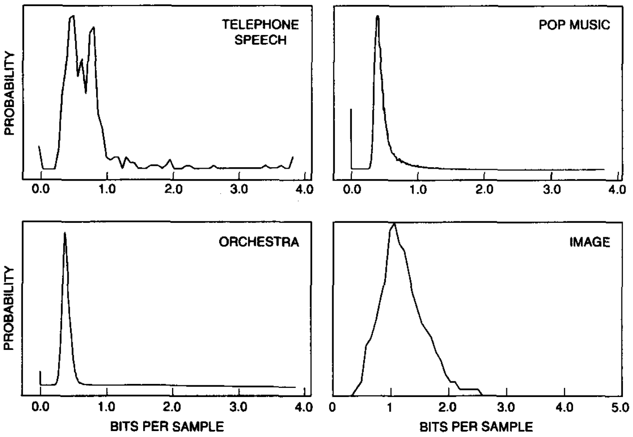


Fig. 24. Histograms of perceptual entropy.

Fig. 5 indeed exceed 2 b/sample, as suggested by the tails in Fig. 24).

Several, perhaps most, coders used for compressing audiovisual information can be regarded as special cases of the perceptual coder of Fig. 23. For example, elimination of the first two boxes leads to well-known classic algorithms. The most degenerate case of Fig. 23 is a memoryless PCM coder.

VI. Perceptual Coding of Speech

As a class, speech compression algorithms have operated below the level of transparency because of technology needs in speech communication. We will treat the historically (and also technically) distinguished cases of narrowband telephone speech (300–3400 Hz) and wideband speech (100–7500 Hz) separately, and we shall examine how perceptual noise shaping principles have been incorporated in respective algorithms to maximize speech quality at a specified bit rate.

1) Telephone Speech: At 64 kbps (8 b/sample at 8 kHz), the use of a nonuniform (logarithmic) 8-b quantizer provides transparent coding of telephone speech. There is very little noise shaping in this process, and in the absence of such a perceptual mechanism, the coder realizes transparency by what may be considered a memoryless, overcoding process with an unweighted SNR value of about 38 dB [65].

With a simple adaptive predictor and an associated mild form of noise shaping (signal-adaptive noise envelope and a lowpass noise spectrum), adaptive differential PCM (ADPCM) provides high-quality speech at 32 kbps (4 b/sample) [65]. The speech quality is inferior to that of 64-kbps PCM in careful listening, although the telephone handset receiver tends to minimize the difference. If the same quality is maintained in going from 8 to 4 b/sample, the SNR gain needs to be 24 dB using a 6-dB-per-bit rule of thumb. In reality, given the slightly lower quality in the 32-kbps coder, the SNR gain may be closer to 18 dB of which about 10 dB can be attributed to a prediction gain (redundancy reduction), and about 8 dB relates to a noise shaping advantage (perceptual coding).

For rates of 16 kbps and lower, both of the above functions need to be addressed more efficiently. Adaptive predictors are made more powerful by the use of formal LPC models and by including pitch prediction, and noise shaping is made more deliberate (than in ADPCM) by relating it to distortion masking by the formant frequencies (vocal tract resonances) in voiced speech. The code-excited linear prediction (CELP) algorithm [5], [20], [44], [72], [139] is a generic structure for efficiently realizing these functions.

The CELP diagram in Fig. 25 is a closed-loop LPC coder. Adaptive prediction (or strictly, an inverse function) is provided by the long- and short-delay correlation filters. The function of these LPC filters (which simulate the vocal apparatus) is to create a speech-like signal from an *innovations* signal, an excitation waveform. In vocoding, the excitation waveform is limited to either a periodic or a noisy signal, in the interest of a very low bit rate. In contrast, the CELP coder seeks to enhance naturalness in the output speech by permitting one of a large set of codevectors as the appropriate excitation for a given speech block, and its LPC description. In the example of Fig. 25, the selected codevector is one out of a codebook of 1024 entries, prechosen either as appropriate noise-like signals (as in a stochastic codebook) or as speech-derived signals (as in an iteratively designed deterministic codebook). The selection of the best codevector is based on minimizing the *perceptually weighted* distance (distortion) between the synthesized speech (output of the LPC filter) and the speech to be coded.

Figure 26 describes the distortion-weighting process. The hearing model is one which claims that the optimal distortion (noise) is neither white noise nor a spectrum parallel to the input speech spectrum (the broken spectra in the figure), but an intermediate (solid-line) characteristic that has a speech-like, *half-white* spectrum. The claim is that, for the same distortion power (area under the curve in

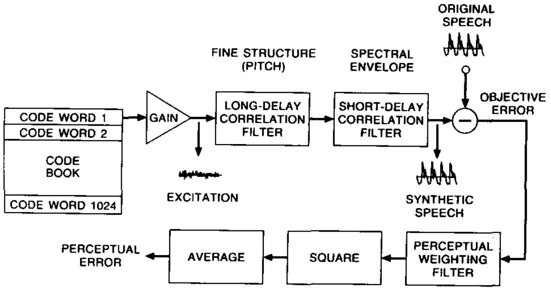


Fig. 25. Code-excited linear prediction (CELP) algorithm for speech coding (after [5]).

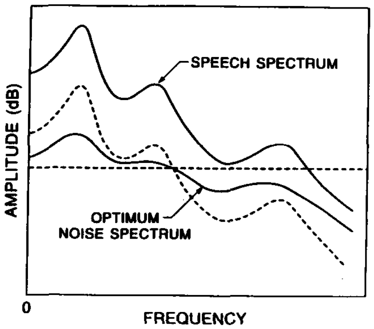


Fig. 26. Three generic forms of noise shaping in the coding of voiced speech (after [6], [65]).

the figure), such a half-white spectrum provides a maximal overall exploitation of distortion masking by the powerful formants in speech. Two observations from Fig. 26 can make this claim a little clearer. First, the white-noise spectrum is suboptimal in the sense that the SNR in the frequency region of the third formant would be too small to mask the corresponding distortion in that high-frequency region. Second, while the speech-like spectrum increases the SNR at the third formant, it does so at the expense of the SNR near the first formant, thereby compromising the low-frequency fidelity of the coder. The intermediate (optimum) distortion spectrum strikes a balance at all frequencies without compromising the very important low-frequency area. This spectrum is in a sense an MND profile, in the language of Fig. 23.

The MND profile of Fig. 26 is a smooth spectral envelope, with the practical advantage that it can be related easily to the speech spectral envelope implied in the LPC coder. In particular, there is very little computational overhead in deriving the optimal noise shape, and no bit-rate overhead needed to perform the perceptual shaping.

What this MND profile lacks, however, is a fine frequency structure related to the pitch frequency of the speaker. If the analysis procedure includes such a fine structure, it is possible to use even finer noise shaping than that used in Fig. 26. This subject will be addressed carefully in the discussion of perceptual audio coding (Section VII).

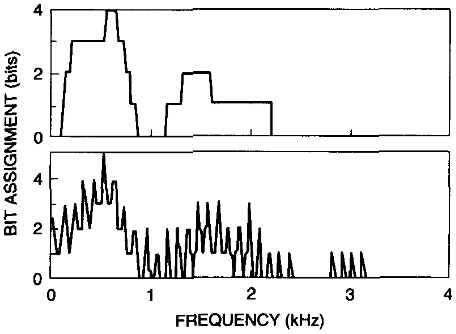


Fig. 27. Adaptive bit allocation based on (a) coarse and (b) fine spectral analysis (after [173] and [65]).

For now, the benefits of finer frequency analysis can be illustrated by a simple example (Fig. 27). This example is not from CELP coding, but rather from adaptive transform coding, but it makes the point. The plots in the figure are bit-allocation profiles versus frequency. The bit-allocation algorithm (and the corresponding noise spectral shape) is driven by a smooth spectrum model in the upper figure and by a finer spectral analysis that includes pitch information in the lower figure. The total number of bits is the same in both examples. In the upper picture, the algorithm responds to average spectral energy and allocates all bits to the low frequencies of speech. It ends up overcoding these frequencies at the expense of entirely losing frequencies above 2.3 kHz (the third formant area of Fig. 26). With the finer analysis in the lower figure, the algorithm recognizes that the pitch-structure-induced *valleys* at low frequencies can in fact be lower in energy than the pitch *peaks* at higher frequencies, and uses this information to allocate quantization bits to selected high frequencies (in particular, the four 1-b spikes centered on 3 kHz). In other words, by responding to a fine, local energy cue, it preserves essential high frequencies in the speech signals. As described here, the algorithm is still steered by energy cues rather than by formal noise-masking principles. But the advantages of a finer frequency analysis carries over to a noise-masking argument as well, and reinforces the general need for a powerful, local signal analyzer (the first box of the perceptual algorithm of Fig. 23).

2) Wideband Speech: Table 4 describes the *Articulation Index* [38], an intelligibility measure, as a function of frequency content in speech. It shows that frequencies below 200 Hz contribute nothing to intelligibility (although such low frequencies improve naturalness and presence). The table also shows that the contribution to articulation due to frequencies above the conventional telephone limit of 3400 Hz is significant. This is true although the contribution to speech energy due to these high frequencies may be smaller, as shown in Fig. 28(a) (note the decibel scale on the vertical axis). In some examples, as in the vocal music illustration of Fig. 28(b), the out-of-telephone-band contribution can be significant even from an energy standpoint. Wideband speech (100–7500 Hz, nominally 0–8 kHz) is difficult to compress in the sense that the spectral dynamic range is high, and expectations of quality tends to be higher

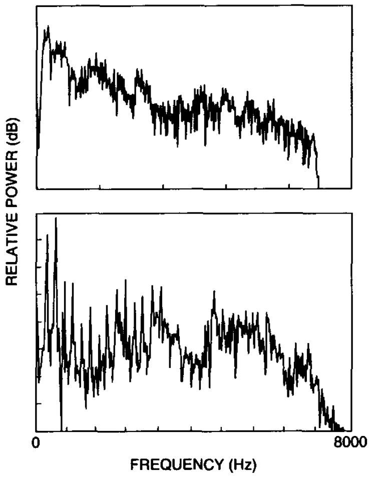


Fig. 28. Spectra of wideband signals in (a) speech and (b) vocal music.

Table 4 Frequency Bands Contributing Equally to the Articulation Index

Number	Limits	Mean	Number	Limits	Mean
1	200 to 330	270	11	1660 to 1830	1740
2	330 to 430	380	12	1830 to 2020	1920
3	430 to 560	490	13	2020 to 2240	2130
4	560 to 700	630	14	2240 to 2500	2370
5	700 to 840	770	15	2500 to 2820	2660
6	840 to 1000	920	16	2820 to 3200	3000
7	1000 to 1150	1070	17	3200 to 3650	3400
8	1150 to 1310	1230	18	3650 to 4250	3950
9	1310 to 1480	1400	19	4250 to 5050	4650
10	1480 to 1660	1570	20	5050 to 6100	5600

in wideband audioconferencing or loudspeaker telephony than in telephone speech communication with a 3.4-kHz bandwidth.

We will now describe a perceptual noise shaping algorithm that has proved to be quite effective for compressing wideband speech to bit rates on the order of 32 to 16 kbps (2 to 1 b/sample, assuming 16-kHz sampling).

The general form of the noise-shaping filter in CELP coding is a weighting function

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)}, \qquad 1 \leq \gamma_2 < \gamma_1 \leq 1$$

where $A(z)$ is the LPC polynomial. The effect of γ_1 or γ_2 is to move the roots of $A(z)$ towards the origin, deemphasizing the spectral peaks of $1/A(z)$. With γ_1 and γ_2 as in the equation, the response of $W(z)$ has valleys (antiformants) at the formant locations and the interformant areas are emphasized. In addition, the amount of an overall spectral roll-off is reduced, compared to the speech spectral envelope as given by $1/A(z)$. Noise is less audible if it shares the same spectral band with a high-level tone-like signal.

The spectral dynamic range of wideband speech is considerably higher than that of telephone speech, and the amplitudes of the 3400- to 7000-Hz components are usually near the bottom of this dynamic range. The unweighted SNR in CELP coding tends to be negative at high frequencies. The auditory system is quite sensitive in this region and the quantization distortions are clearly audible in a form of crackling and hiss. Noise weighting is therefore very crucial in wideband CELP and the balance of low- and high-frequency fidelity is quite delicate.

The filter $W(z)$ in the above equation has an inherent limitation in modeling the formant structure and spectral tilt concurrently. The spectral tilt is more or less controlled by the difference $\gamma_1 - \gamma_2$. The tilt is global in nature and it is not possible to emphasize it separately at high frequencies. Also, changing the tilt affects the shape of the formants of $W(z)$. A pronounced tilt is obtained along with higher and wider formants, which puts too much noise at low frequencies and in between the formants. The formant and tilt problems need to be *decoupled.* The approach is to use $W(z)$ only for formant modeling and to add another section for controlling the tilt only [64]. The general form of the new filter is

$$W'(z) = W(z)P(z)$$

where $P(z)$ is responsible for the tilt only. Various forms of $P(z)$ have been studied, and the following two-pole section has been proposed based on listening tests:

$$P(z) = \frac{1}{1 + \sum_{i=1}^{2} p_i \gamma_p^i z^{-i}}.$$

The coefficients p_i are found by applying the standard LPC algorithm to the first three correlation coefficients of the impulse response of the current-frame LPC inverse filter. The parameter γ_p is used to adjust the spectral tilt of $P(z)$. The values $\gamma_p = 0.7$, $\gamma_1 = 0.95$ and $\gamma_2 = 0.8$ have been found to yield the best perceptual performance.

Figure 29 demonstrates the effect of the enhanced noise-shaping filter. The broken curve in the figure shows a typical spectrum of a conventional inverse filter $W^{-1}(z)$. The solid curve is the spectrum of an enhanced inverse filter $W^{-1}(z)P^{-1}(z)$ for the same underlying LPC filter. For the same general tilt, the enhanced filter has less pronounced formants especially for the lowest and highest formants. Compared to the conventional filter, the enhanced filter

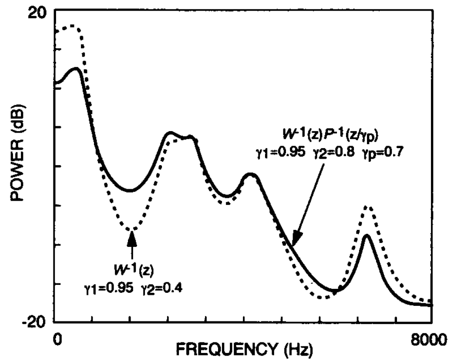


Fig. 29. Noise shaping in the coding of wideband speech (after [64]).

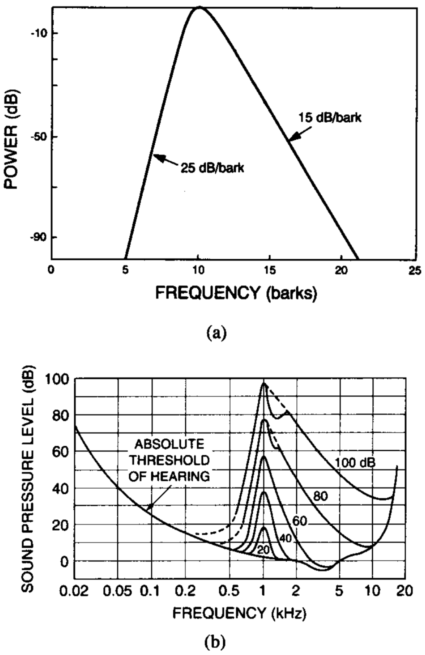
attenuates the noise by about 5 dB at the lowest and highest formants, while achieving the right overall spectral tilt.

The CELP coder with the noise weighting described above has been implemented with an LPC predictor order of 32, a coding block size of 5, a codebook size of 1024, and without a pitch loop. The performance of the computer-simulated 32-kbps CELP coder has been assessed by comparing it to the G.722 CCITT standard wideband coder at 64 kbps [16]. The test material included 4 male and 4 female utterances. Each utterance was coded by the G.722 algorithm and by the CELP to form a pair of utterances. The order in a pair was set at random. However, each pair was played in both possible orders to eliminate biased decisions. Twenty listeners took part in the test. Each listener was asked to vote for the better sounding utterance or to split the vote equally, if no preference could be made. The final scores were defined as the percent number of votes for each system per given condition.

The overall preference scores were 51.72% and 48.28% for CELP and G.722, respectively [64]. This means that, on average, the two systems performed alike, which is extremely encouraging, recalling that the CELP bit rate is half that of G.722. Finally, the coding delay of the G.722 coder is 1.5 ms whereas that of the CELP coder is only 0.94 ms. The results of the experiment also suggest that the *mos* quality score of the 32-kbps audio coder is close to 4.0. Extensions of noise-shaping algorithms to 16 kbps have produced promising results [40], [127], [166], although the quality is noticeably lower than that at 32 kbps.

VII. Perceptual Coding of Wideband Audio

The term *masking* describes the effect by which a fainter but distinctly audible signal becomes inaudible when a correspondingly louder signal occurs approximately simultaneously. Earlier results on the *tone-masking-noise* [132] problem showed that there is a band of frequencies centered at each tone in which the just-noticeable noise energy remains nearly constant, even though the noise bandwidth and/or in-band noise shape are changed. This band of frequencies was called the *critical band* (the staircase treads in Fig. 31). Later research on the *noise-masking-tone* problem [56] showed that noise maskers had the same sort of behavior and that the critical band corresponded to a physical measurement in the cochlea.



(a)

(b)

Fig. 30. (a) The spreading function for auditory noise masking. (b) The elevation of auditory threshold due to a critical-band noise masker of varying intensity at 1 kHz.

Given this critical band scale, masking of steady-state tones and noise inside of one critical band is relatively well described. In the course of investigating masking phenomena outside a critical band, the *spreading function* (Fig. 30(a)) [135] was postulated; this describes, for steady-state situations, the masking effect of a signal in one critical band on signals in another critical band. This spreading function is currently believed to be a by-product of the mechanical cochlear filtering mechanism. The shape of the spreading function varies with level, and the masking abilities of a signal spread farther from the base frequency as the level of the masker in increased. Note also that Fig. 30(a) suggests that for a given frequency separation, a masker does a better job of masking a higher frequency maskee than a lower-frequency maskee: a phenomenon sometimes called the *upward spread of masking*.

Figure 30(b) refers to an early experiment in this subject. It illustrates the level of a test tone just masked by critical-band wide noise with center frequency of 1 kHz and different levels, as a function of the frequency of the test tone [174], [175]. The figure includes the absolute threshold of hearing as a baseline. The shapes of the various elevated characteristics are similar to the spreading function in Fig. 30(a). The exact shapes, however, are dependent on the noise level. The double peaks in the two uppermost characteristics are due to nonlinear phenomena in hearing, resulting in a second peak that approaches the second harmonic (2 kHz) at high loudness levels. The concept of noise-masking tone is a significant departure from models used in earlier coding theory. There, it was usual to test the coder with a tone input and model the coding distortion as noise. In adaptive perceptual coding, we recognize that the signal in a critical band can be noise-like, while a distortion component could be very localized and tone-like; hence the noise-masking-tone model of Fig. 30(b). Results for *tone-masking noise* tend to be more complicated.

The results from masking experiments are sometimes summarized by the equations below where E_T and E_N are tone and noise energies, B is the critical band number, and the left-hand sides of the equations represent the maximum energy of the masked signal:

$$\begin{aligned} &\textit{Tone Masking Noise:} \quad E_N = E_T - 14.5 - B \text{ (dB)} \\ &\textit{Noise Masking Tone:} \quad E_T = E_N - K \text{ (dB)} \end{aligned}$$

Various values in the range of 3 to 6 dB have been proposed by experimenters for the parameter K. The variability has to do with differences in experimental paradigms and differences in the way intersubject variations are accounted for. In Fig. 30(b), K is close to 3 dB. A value of 4 dB is proposed in [56]. A value of 5 dB is used in some of the more recent work reported in this paper [69].

Figure 31 shows a JND profile for the example of a trumpet sound as input signal. The JND function in Fig. 31 is based on an interpolation of tone-masking noise and noise-masking tone results, followed by the application of the spreading functions. This sort of interpolation is necessary because the classical results are for pure tones, or for critical-band noise, and not for mixed signals such as speech and music. In practice, such interpolations, coupled with the critical band model, provide a very good estimate of the masking threshold for most signals. In sophisticated versions of the algorithm, the interpolation is done on a frequency-by-frequency basis, using a measurement of the "tonality" at each frequency [68], [69]. These interpolated noise-masking-tone and tone-masking-noise thresholds are used to calculate a level below that of the spread energy in the short-term signal. This is the level below which injected noise will be masked. A version of this method can be found as "Psychoacoustic Model II," in the ISO-MPEG-1 audio coder [108].

Figure 31 illustrates a generic and powerful way of incorporating a human auditory model in the coding process [11], [67], [68], [69], [146]. The method is to frequency-analyze the input audio signal and to postulate a just noticeable quantization noise threshold as a function of frequency for the given audio signal spectrum. Quantization errors that are equal to, or lower than, this threshold are masked by the audio spectrum in the frequency vicinity of the errors. Given the noise threshold, a variable bit allocation procedure can be utilized to provide an actual noise spectrum close to the threshold characteristic. Extremely high compression efficiency can be realized because the sophisticated noise threshold typically permits very coarse, or even 0-b, quantization of significant fractions of the input spectrum. The implementation of the perceptual coder may use two interacting processes: a function that implements the JND, or an estimate of the JND, including a controllable parameter C that causes overcoding or undercoding; and a bit-counting function that counts or estimates the bit rate resulting from the quantizing and entropy-coding systems. In one example of an iterative loop, the difference between the actual and target bit rates in a given stage of iteration causes a change in the parameter C. Several cycles of iteration result in a stable combination of desired bit rate and the approximate JND or MND condition that is realizable at that bit rate. The final bit allocation as a function of frequency, together with stepsizes for the various quantizers, are then coded and transmitted to the receiver.

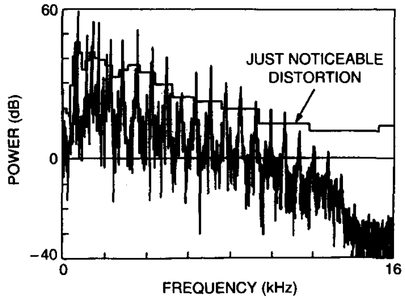


Fig. 31. Just-noticeable distortion (JND) versus frequency for one block of a trumpet signal.

The objective signal-to-noise ratio in a perceptual audio coder is typically fairly low (for example, as low as 20 dB), while the subjective signal-to-noise ratio is at a level high enough to signify transparency or near-transparency. In one example, with a signal which consists of a male *a capella* chorus, the coder is perceptually transparent at a 13-dB overall SNR, when the noise is inserted according to psychoacoustic principles. On the other hand, the coder is nontransparent to additive white noise at an SNR as high as 60 dB, and nontransparent to locally adjusted white noise at an SNR of 45 dB. With a 1-kHz tone input, transparency with additive white noise requires an SNR of 90 dB. With ideally shaped noise, transparency occurs at an SNR of 25 dB.

The JND function is a function of a finely described input spectrum and the ear model, rather than a simple transformation of the LPC spectrum of the input signal, as in Figs. 26 and 29. While it is conceivable that a very-high-order LPC analysis can describe the input spectrum finely enough to permit a useful JND model, current methodologies for transparent or near-transparent audio coding have depended on high-resolution frequency analysis using either a subband coding or transform coding framework. Currently, high-order LPC methods have also been limited by their ability to adapt to a changing masking threshold and their ability to provide a predictable trajectory for adaptation.

1) Simultaneous (Frequency) and Nonsimultaneous (Temporal) Masking: The distortion masking implied in Fig. 31 is a frequency-domain effect. There is also evidence of temporal masking in the human auditory mechanism. A

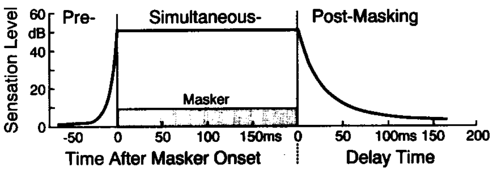


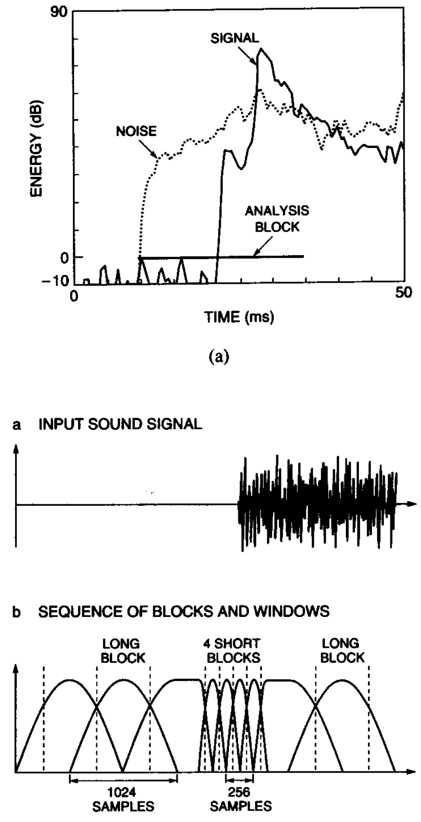
Fig. 32. Temporal masking in the human auditory system.

strong signal can mask a weaker signal that occurs *after* it (*post-masking* or *forward masking*), and a weaker signal that occurs *before* it (*pre-masking* or *backward-masking*). This is shown in Fig. 32 [166], [175]. Note that in this figure, post-masking uses a different time origin than pre-masking and simultaneous (frequency-domain) masking. The main result of Fig. 32 is that forward masking lasts longer than backward masking. But in both cases, the skirt of the masking characteristic is quite sharp. Forward masking is easier to realize in the sense that there is no need to predict the availability or onset of a maskee, as in backward masking.

One of the potential degradations in low-bit-rate transform coding of audio is the phenomenon of *pre-echo*. Consider the coding of a sound signal characterized by a sharp attack. Consider further the situation where the signal begins only in the ending part of the transform block. The inverse transform has the effect of spreading the quantizing distortion throughout the transform block. The result is a block in which the distortion extends through the block, while the signal appears only at the end of it: hence the pre-echo effect (Fig. 33(a)). This effect can be mitigated by backward masking, but not if the block length is very long. As in the application of masking thresholds, it is again wise to use a pre-echo criterion matched to the critical listener, so that one can detect and eliminate pre-echo, for high-quality coding.

A natural technique to minimize pre-echo distortion is to use a short transform block length. However, in order to ensure the coding gains associated with a long transform block, long block lengths should be used when a sharp onset is not likely, as in steady-state audio segments. The result is a variable-block length transform coder (Fig. 33(b)). The block length in such a system may have typically two or four values, for example: 128, 256, 512, and 1024 samples (at 48 kHz). In the illustration of Fig. 33(b), two block lengths are used, together with asymmetrical start and stop windows while making block-length changes. One of the challenges in designing such a coder is the use of an effective criterion for block-length adaptation. Perceptual entropy gradients have been shown to work better in this connection than simple energy gradients.

2) The ISO-MPEG Audio Coding Standard: This is a three-layer standard where audio quality increases monotonically with increasing layer number [60], [98], [108]. The first two layers are based on a polyphase filterbank with 32 equally spaced bands, each with a bandwidth of 750 Hz and a sampling rate of 1.5 kHz. Decimated subband waveforms are processed in 8-ms blocks for adaptive quantization. Layer 2 is characterized by even more efficient algorithms for quantization and perceptual coding. The third layer is based on finer frequency analysis obtained by an MDCT-transform decomposition of the subbands in earlier layers. It also provides advanced pre-echo control, adaptive entropy coding and the iterative loops mentioned earlier for optimum coding.



(a)

(b)

Fig. 33. (a) Spreading of noise in time: input signal (attack of a triangle), solid line; analysis block in a transform coder; and noise produced by the transform coder, broken line. (b) Transforms with two different block lengths.

Figure 34 illustrates a *hybrid* filterbank composed of a subband stage followed by a transform coding stage [68]. Here, as in the ISO-MPEG standard, the finer frequency resolution provided by the MDCT stages is necessary for maximum utilization of distortion-masking effects. The finer frequency resolution, and the longer time window associated with it, are also critical for realizing powerful forms of joint stereo coding. The hybrid filterbank is shown to result in lower estimates of perceptual entropy for many, but not all audio inputs [68]. An input-dependent filterbank, while admittedly very complex, would result in consistent and substantial reductions of perceptual entropy.

3) Joint Stereo Coding: One mode of the ISO standard codes the left and right channels of a stereo pair as independent sound channels. There is also a provision for jointly coding the stereo pair, but the filterbank and bit

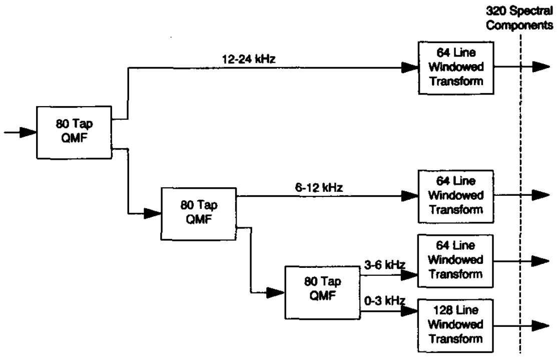

320 Spectral Components
12-24 kHz
64 Line Windowed Transform
80 Tap QMF
80 Tap QMF
6-12 kHz
64 Line Windowed Transform
64 Line Windowed Transform
3-6 kHz
80 Tap QMF
0-3 kHz
128 Line Windowed Transform

Fig. 34. A hybrid subband-transform analysis system for audio [68].

stream specifications in the ISO standard are such that the full potential of joint stereo coding is not realized in the ISO algorithm.

The two channels of a stereo pair are in general correlated and, as expected, there is a significant advantage in coding these channels jointly, as in Fig. 35. In a classical coding approach to the problem, the left (L) and right (R) channels are transformed into a sum ($L + R$) and difference ($L - R$) pair, and unequal bit allocation based on energy-related metrics will provide an overall minimum mse. In perceptual coding, the intention is to effect maximal masking, while making efficient use of signal redundancies. When operating on a stereo signal, there are several intricate issues [69]. One of these is minimizing the unmasking of noise as related to *masking level differences* (mld). When a low-frequency signal (2 kHz or below) is received by both ears, the phase relationships in the signal and any independently injected noise may result in up to 15 dB of *binaural unmasking*. This means that the R and L coders may mask the R and L distortions completely, while the binaural signal may have significant unmasked noise. In the case where a strong central or antiphase signal exists, the *mld* can be controlled by the use of sum/difference coding, a situation that also extracts the redundancy present in the stereo signal. Another consideration is that the stereo coder must not create a situation where sum/difference coding is less efficient than L/R coding. If the masking thresholds for L and R are substantially different, any coding in the sum/difference domain will have to be at the lower of the two thresholds. This can result in substantial inefficiency in coding the channel with the higher threshold, unless L/R coding is used. In general, this difference in L/R threshold also shows the existence of a situation where L and R are not strongly correlated in the frequency domain. A perceptually efficient joint stereo technique looks at the difference between the L and R perceptual thresholds, and chooses the $(L + R, L - R)$ decomposition *only* if that difference is smaller than a carefully selected threshold. The switch between the stereo-coding modes occurs both in time and in frequency [69].

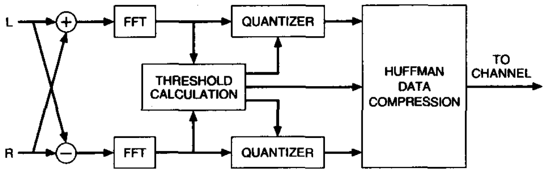


Fig. 35. Block diagram of joint stereo coding.

Figure 36 displays a set of eight spectra related to joint stereo coding. The input is a block of vocal music (the same as the input used in Figure 28b). An interesting effect in the figure is that the sum-threshold (sum-JND) and difference-threshold (difference-JND) characteristics tend to be close to each other even though the sum and difference spectra are very different. As a result, we are able to discard a good deal of the difference component in the coding process, without introducing any perceived noise, either in the individual L and R channels, or due to the binaural *mld*.

In the particular example of the vocal signal in Fig. 28(b) (and Fig. 36), it turns out that $(L + R, L - R)$ coding is used almost all of the time except during silences.

VIII. Perceptual Image Coding

Over the years, there has been a steady increase in the extent to which knowledge about human perception has been incorporated into image compression. Most of the early work has utilized the *frequency sensitivity* of the human visual system as described by the *modulation transfer function* (MTF). This function describes the sensitivity of the human eye to sine-wave gratings at various frequencies. The psychovisual experiments in [17], [23], and [92] have proposed a commonly used model for the MTF. Since the MTF is defined for sine-wave inputs, it is not directly usable for DCT-based coders. Recent work [106] has developed a transformation that accounts for the difference between the Fourier and DCT bases. Similar procedures could be used to transform the model to other frameworks for frequency analysis. From these models, given that the minimum viewing distance is fixed, it is possible to determine a static JND threshold for each frequency band. These thresholds can be used for both quantization [156] and bit allocation [114].

The models can be extended to include *contrast sensitivity* as well. The homomorphic models mentioned earlier also attempt to utilize these properties. Contrast sensitivity can be included either implicitly, by developing functions of masking threshold versus local brightness, or explicitly, by using homomorphic systems or so-called *perceptually uniform* or *equi-luminant color spaces* [150].

The main problem with the aforementioned systems is that they do not go far enough in utilizing the masking properties of the human visual system. Frequency sensitivity is a global property dependent only on the image size and viewing conditions. Contrast sensitivity has been exploited mainly via pre- and postprocessing as in the homomorphic models. What is needed is a more dynamic model that allows for finer control of the quantization process.

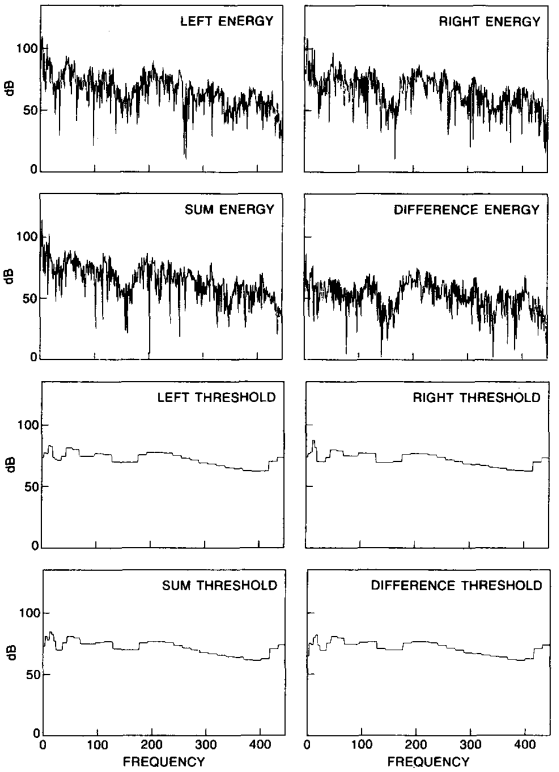


Fig. 36. Signal and JND spectra in joint stereo coding.

Analogous to the frequency-dependent masking functions in audition, there is evidence for similar processes in the visual system [144]. These properties depend on the local scene content and therefore have the desired property of *local control.* In other words, the local JND level depends on the MTF of the human visual system, the local contrast sensitivity, and a measure of the local texture. In order to achieve the highest subjective quality at a given bit rate, all of these properties must be utilized.

Several coders have been proposed that utilize these properties to achieve higher perceived quality for a given bit rate. The work in [21] utilizes the concept of a fixed quantization matrix, which accounts for the MTF of the visual system, in conjunction with the concept of encoding only the portion of the signal that exceeds a local threshold. This produces a coder structure that is the predecessor of contemporary perceptual coders. More recent systems [22], [95], [105] utilize frequency and contrast sensitivity, and in addition use some form of texture masking. In these methods, instead of explicitly computing JND thresholds incorporating frequency, contrast, and texture, each DCT block is classified according to its energy. Based on this classification, an appropriate quantization rule is invoked.

As with perceptual audio compression, the choice of the filterbank used in a coder can affect the performance of the compression system. The primary purpose of perceptual quantization is to control the spatial frequency location of the quantization distortion. Therefore, the filterbank should be selected to allow this form of control. Ideally, the addition of quantization distortion to one coefficient should not show up in coefficients that are not adjacent to the one that was perturbed. In addition, the analysis filterbank should mimic the structure of the visual system. For a human, this structure is a set of filters with frequency spacing of about 1.5 octaves and an angular width of about 40°.

The DCT (and other uniform filterbanks) meet the first criterion [78], but do not meet the second. This leads to difficulty in creating masking models since there is a mismatch between the underlying structure of the model and the structure of the transform that it is being implemented. One means of implementing a transform that mimics the visual system is the *Cortex Transform* [161], and this results in an image coder based on that transform [162]. The main stumbling block in using this approach is that the Cortex Transform is not a maximally decimated filterbank. This results in an expansion in the amount of data that need to be encoded. At present, visual masking models have not been able to provide enough coding gain to overcome this disadvantage.

As pointed out earlier, signal-to-noise ratio is not an accurate predictor of subjective image quality, especially at low bit rates [2], [13], [80], [86], [100], [126]. One way of addressing this problem is to transform the errors into a space where mean square error could be used. This is the approach used in Fig. 20 as well as in [106] and [130]. The original and coded images are transformed into a "human visual space" by utilizing Weber's Law and the MTF. This transformation is designed to compensate for the effect of those portions of the visual system. The mean squared error between the transformed images provides a weighted distortion, as in Fig. 20. This approach produces better correlation to subjective evaluation than the unweighted *mse*, but is limited by the fact that the visual model is global. To alleviate this problem, the concept of a visual difference predictor is introduced in [30]. A detailed visual model is applied to the original image and local JND thresholds are computed for each pixel in the image. The differences between the original and compressed images are then computed and transformed into JND units. The resulting representation gives a good representation of image quality. Portions of the input that have errors on the magnitude of a JND show little or no visible degradation, while those portions with larger amplitudes correlate well with the portions where coding impairments are visible.

As was shown in Section VII, a more efficient coder can be generated by utilizing a good local perceptual model to identify the irrelevant portions of a signal working in concert with an efficient redundancy removal system. This approach of structuring a coder around generating an MND profile and then efficiently quantizing the resulting information [80], [128], [129] will now be presented.

1) The JND-Based Image Coder: The starting point for the locally adaptive perceptual coder is the two-dimensional subband decomposition of Fig. 12(b). Image blocks in each of 16 subbands are further analyzed to determine local

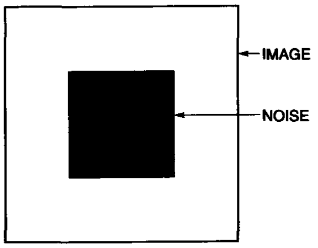


Fig. 37. An experiment to evaluate the visibility of image distortion.

Table 5 Just Noticeable Noise Levels in 2D Subband Coding with 16 Subbands (Entries are rms values, for constant mid-gray background level of 127; rows and columns represent horizontal and vertical frequency.)

0.25	0.40	2.0	6.0
0.50	1.0	4.0	8.0
2.0	3.0	4.0	6.0
3.0	6.0	10.0	11.0

values of brightness and texture (as in the first block of Fig. 23).

Distortion sensitivity profiles are then derived as functions of frequency (subband number), brightness, and texture [128] and color parameters [129]. This is done by using the basic experimental paradigm of Fig. 37 in which a subject views the variable-power noise square at a specified distance and determines the noise variance for which the noise square is *just noticeable* against a specified background of flat texture. (To be precise, the random noise is added to a frequency band in question, and the corrupted images are passed through the synthesis filterbank.) In the cited experiments, the noise had approximately the same mean value as the background, and its density function was uniform. The experiment is repeated at various background brightness levels to determine the *brightness sensitivity* (Fig.38(a)). When the mean value of the noise square is the same as that of the background, the noise square tends to be most visible against a mid-grey background, assuming that the input image includes gamma correction [100]. The *frequency sensitivity* can be inferred from the eye-sensitivity diagram, the MTF (Fig. 38(b), a result of the more general Fig. 11); alternatively, the frequency sensitivity is experimentally measured by repeating the basic subjective experiment for each of the 16 subbands. Table 5 shows the experimentally measured frequency sensitivity for a viewing distance of six times the picture height and for a mid-grey background of flat texture. Finally, the *texture sensitivity* is determined by an empirical calculation in which texture is defined as a local intrablock variance and the MTF is (again) used to establish a relation between visibility of distortion and texture.

The upper part of Fig. 39 depicts a constant texture background with white noise of rms value 8 added to the center third of the lowest subband of the right half of the picture. The figure de monstrates the expected phenomenon that distortion visibility is low when the background has a strong texture. The lower picture in Fig. 39 depicts a constant level of noise (rms value of 8 for a peak input of 255) added to the center third of the lowest subband of a ramp input whose brightness changes continuously from white (at the left) to black (at the right). The figure demonstrates that distortion visibility is greatest against a mid-grey background.

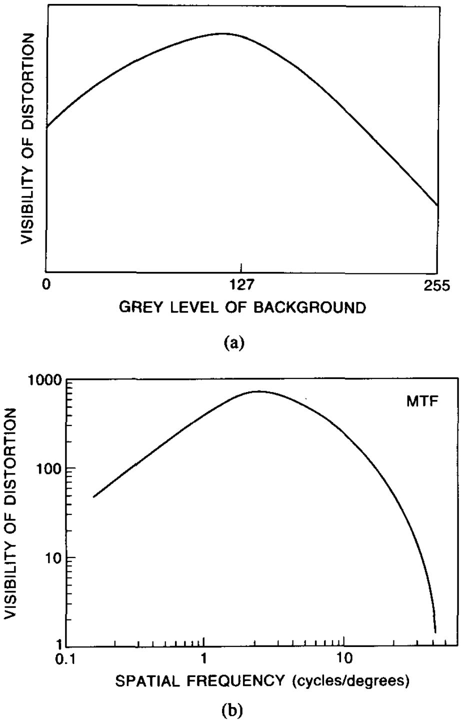


Fig. 38. Distortion visibility as a function of (a) background brightness and (b) spatial frequency.

The interactions between the various parameters controlling noise sensitivity are small. Hence component effects can be superposed to predict overall noise visibility as a function of various input signal parameters.

Data from the above generic experiment are used to derive JND estimates as spatial profiles for each of the 16 subbands of the image to be coded. These JND profiles lead in turn to spatially adaptive bit allocation in each subband (see Fig. 23). With a typical image, several spatial blocks are allocated zero bits, and in fact a large number of subbands see zero-bit allocation to *all* of their spatial blocks.

These effects are illustrated in Fig. 40. Part (a) of the figure is a 512 × 512 pixel image of a head-and-shoulders image. Part (b) is the image of the first subband (lowest horizontal and vertical frequencies). Part (c) is the JND image of part (b): dark areas in (c) represent parts of (a) where the JND is high, implying the possibility of coarse quantization; white areas in (c) represent parts of (a) where the JND is low, and intermediate intensities in (c) signify intermediate values of JND. Corresponding results

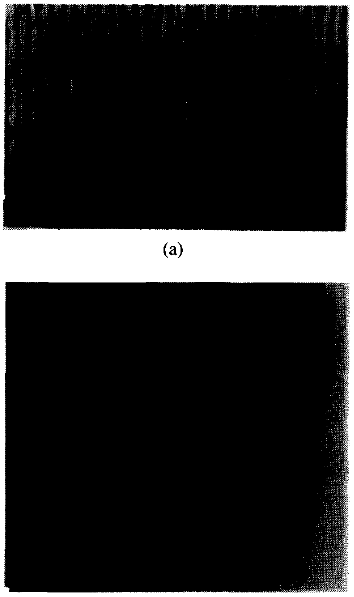


Fig. 39. Illustration of distortion masking by (a) background texture and (b) background brightness.

of variable bit allocation (including zero-bit allocation) are shown in part (d): white blocks in (d) signify that for corresponding regions in the full-band input (a), 4 out of 16 subbands are retained; light-grey blocks in (d) signify the retention of 3 subbands; dark-grey blocks in (d) signify the retention of 2 subbands; and black blocks in (d) signify that only 1 out of 16 subbands (almost always the lowest frequency subband) is retained for corresponding input blocks. As a result of the JND paradigm, the input (a) can be compressed in a perceptually lossless fashion at a bit rate of about 0.7 b/pixel assuming a viewing distance of six times the picture height.

Figure 41 shows the result of perceptual subband coding of a 24-b color input (704 × 480 pixels) at rates of 1.0, 0.67, and 0.33 b/pixel. Distortion is noticeable only at the very lowest bit rate. On a 64-kbps transmission line, the time needed to transmit the 24-b original is about 100 s. With 0.33 b/pixel coding, the transmission time is 1.3 s.

2) The ISO-JPEG Image Coding Standard: The JPEG system [156] is based on DCT coding rather than subband coding. Adaptive methods for bit allocation and quantization provide a very good tradeoff between quality and bit rate. However, with various informally tested images, the perceptual subband coder provides higher image quality than the JPEG algorithm at any given bit rate, and thus achieves transparency at a lower bit rate as well.

The JPEG system [113] is based on 8×8 DCT coding and utilizes a single quantization matrix per image. For the baseline encoder, the input image is divided into 8 × 8 blocks, and then transformed using the DCT. Next, the transform coefficients are quantized using a user-specifiable quantization matrix. The quantized coefficients are then noiselessly coded using either arithmetic or Huffman coding.

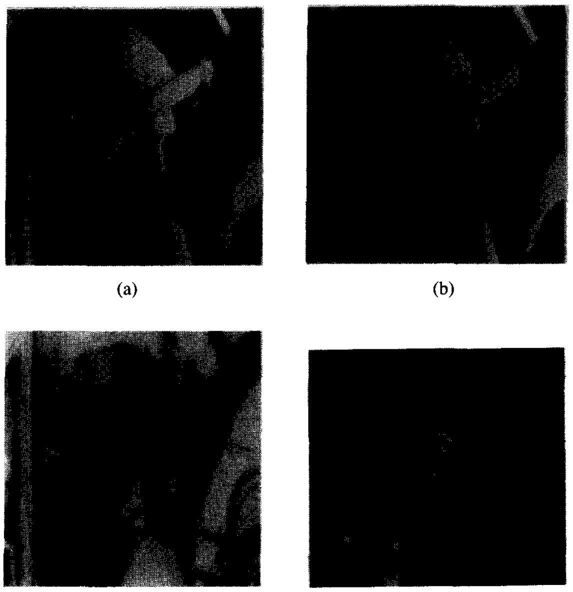


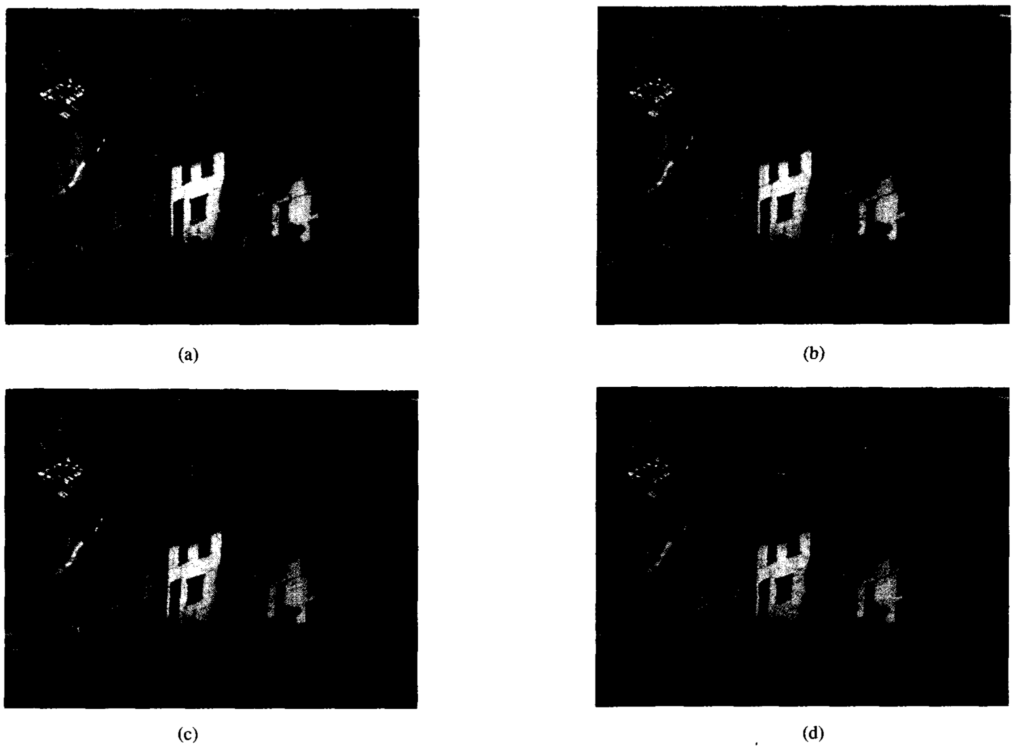
Fig. 40. Perceptual sub-band coding. (a) 512 × 512 input, (b) lowest-frequency subband, (c) spatial JND profile for (b), and (d) spatial profile of the number of subbands retained (out of 16).

Given this structure, it is possible to design quantization matrices that take into account the MTF of the human visual system. Recent work [3], [115] has shown how to derive the perceptually optimum quantization matrices for a given viewing distance for various color spaces. These matrices do not take into account texture sensitivity and may not give perceptually uniform distortion for supra-threshold conditions. The work in [163] provides a design procedure to compute an image-dependent quantization matrix that takes these additional factors into account. It allows the user to specify the average desired distortion in units of JND, and then determines a quantization matrix that will provide that level of fidelity.

These approaches produce a coder that provide subjectively better quality images than using the matrices provided in the JPEG standard. However, they can only work on a global level since the quantization matrix cannot be changed within an image. It is possible to use more explicit perceptual models and a technique called *pre-quantization* to provide more local adaptivity to the quantization system while still maintaining compatibility with the JPEG standard.

In this approach, the JND data from the subband coding experiment are used to enhance the JPEG encoder while retaining the standardized bit stream syntax and decoder algorithm. This is done by using the perceptual methodology as a preprocessor which performs all zero-bit allocations *before* the usual bit allocation and quantization operations in the JPEG quantizer (Fig. 42).

The perceptual model described in conjunction with the subband coder of the previous section provides a noise visibility threshold for transform coefficients. This model


(a)

(b)

(c)

(d)

Fig. 41. Perceptual sub-band coding of color image. (a) 24-bit original (top left) and compressed image at (b) 1.00 (top right), (c) 0.67 (bottom left) and (d) 0.33 (bottom right) bpp.

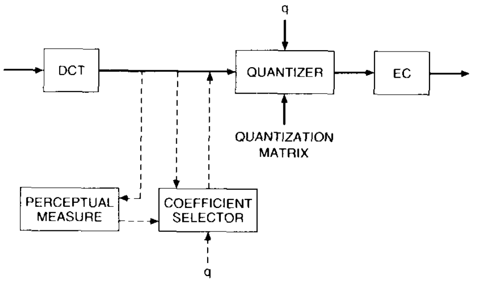


Fig. 42. ISO-JPEG algorithm with perceptual preprocessor.

can be adapted to work with the DCT as well as the subband framework. If a DCT coefficient is less than its corresponding MND threshold, that coefficient can be set to zero before the normal JPEG quantization step.

The result is a picture-dependent gain in the bit rate needed for transparent coding. This gain is on the order of 10 to 45% with typical color inputs. In the example of Fig. 43, for the so-called default quality level specified in the JPEG system, the perceptual preprocessor has the effect of reducing the bit rate from 2.1 to 1.6 b/pixel. The approach of perceptual preprocessing is generic, and can be used with any existing coder.

3) Perceptual Coding of Facsimile: The low bit rates provided by perceptual image coding suggest a possible new approach to the transmission of grey-level facsimile. Rather than transmitting a grey-level photograph by first halftoning it and transmitting it as black–white document of high spatial resolution, one could transmit the picture *as a grey-level image* (using subband coding) and do the halftoning *at the receiver.* For the example of Fig. 44, two alternatives were specifically compared: conventional fax transmission of a 1536 x 1536 half-toned image, using standard black–white coding techniques; and the new method of transmitting a 512 x 512 grey-level image and using halftoning at the receiver. The compression algorithms were arranged to be lossless in both cases, with a perceptual losslessness criterion in the subband method. The grey-level transmission method resulted in a 50% decrease of the bit rate. More significantly, the availability of grey-level information (in the new technique) permits the use of new model-based techniques for efficient halftoning at the receiver. The models used include the HVS model and the model of the printer itself. Fig. 44 illustrates the effect. With the same (300-dot-per-inch) printer, the left image is the result of conventional halftoning. The (sharper) right image is the result of model-based halftoning at the receiver, a precondition for which is the transmission of the picture as a grey-level image [104], [112].

IX. Perceptual Coding of Video

This is an area where the use of perceptual cues are least understood. For instance, we have only begun to use the JND methodology of Fig. 23 for quantizer design. On

Fig. 43. Comparison of JPEG image coding algorithm: without preprocessor (2.1 bpp, lower image) and with perceptual preprocessor (1.6 bpp, upper image).

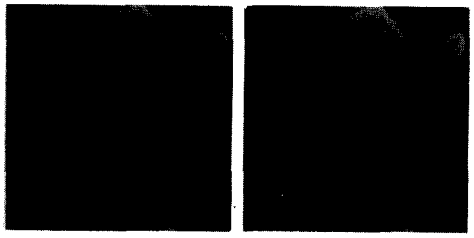


Fig. 44. Grey level transmission of fax document using perceptual coding at 0.5 bpp (left image); and model based halftoning at the receiver (right image) (after [104] and [112]).

a different level, even prior to quantization, we are still using very informal criteria for fundamental designs such as choice of temporal (and spatial) resolutions and the degree of chrominance subsampling.

In the following, we shall comment on two fundamentally different approaches to video coding, and examine the inroads made by perceptual designs in each case.

1) Motion Compensation Followed by 2D DCT: This is a hybrid approach to 3D coding where a differential interframe process is followed by frequency-domain coding of the motion-compensated interframe error (Fig. 45). This

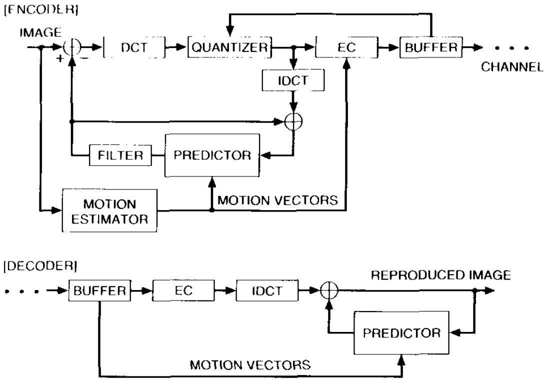


Fig. 45. Block diagram of video coder using motion compensation followed by 2-D DCT and entropy coding (EC).

method is the basis for several international (CCITT and ISO-MPEG) standards for videoconferencing and addressable video [85], [88]. The Phase-1 MPEG standard uses a bit rate of about 1 Mbps; the CCITT coder operates at $p*64$ kbps, with p varying from 1 to 24.

The storage-oriented ISO-MPEG system represents a periodic subset of frames in a high-resolution *intraframe coding* mode (or *intra*mode), in the interest of addressability. In the interframe mode, the coders use motion compensation to generate a prediction of the current video frame from previous (and, in the case of ISO-MPEG, possibly future) frames. The difference between this prediction and the actual frame is then encoded. The frame differences in the CCITT and ISO-MPEG coders are encoded using techniques similar to those in the JPEG coder described previously.

As in JPEG, a global quantization matrix is used, but there is an additional degree of freedom available. On a macroblock level (4 adjacent 8 × 8 luma blocks), it is possible to adjust the scaling of the quantization matrix. This adjustment parameter is introduced to allow the codec to be constant rate, but it can also be used to adjust for local variations in masking level.

One approach to incorporating perceptually based quantization into motion compensated DCT coders is that described in [120]. The approach is to modify the quantizer scaling parameter based on perceptual criteria. First, eight different classes of input-scene complexity are derived and target bit allocations are computed for each class. Then, each macroblock is classified as either *textured* or *low-detail* based on it variance. Next, the textured blocks are further discriminated based on whether the texture is either structured, i.e., *edge-like*, or unstructured. Based on these classifications and the global scene complexity, the quantizer scaling parameter is then set. This differs from the strict JND/MND approach in that perceptual thresholds are not specifically determined, but are implicit in the adjustment rules.

Finally, the prequantization technique described in the Section VIII can also be adapted to MPEG and CCITT systems. The data being encoded are interframe differences,

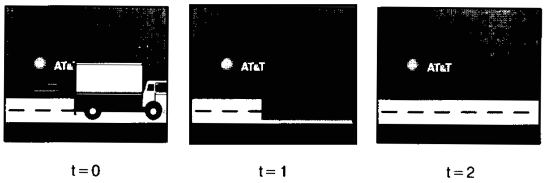


Fig. 46. Explanation of temporal noise masking in video.

and the perceptual model provides MND thresholds. If the local difference is smaller than the MND threshold at that locality, the coefficient does not need to be transmitted. For MPEG-type encoders running in a constant quality mode, a rate reduction of 5–20% is achievable with no visible loss in picture quality.

A low-bit-rate HDTV coder based on perceptual coding of the interframe error has been recently described. The methodology used here was similar to that of Fig. 23, but the extension was approximate in that the structure of the interframe error was very different from the uniform-noise model used in the JND experiment of Fig. 37. Further, the HDTV viewing distance was smaller than the six-times picture height model of that experiment. These effects are offset by the fact that the JND can now be a function of the time dimension as well, and the benefits of temporal masking are now available. Figure 46 explains an example of temporal masking. In the context of newly uncovered background (due to the passing truck), there is a small latency period (say, the frame called $t = 1$) during which the newly exposed areas do not have to be perfectly reconstructed.

HDTV images (1280 × 720 pixels, 60 frames/s) are compressed to 17 Mbps using the above procedure, to facilitate simulcasting over 6-MHz NTSC channels [102]. The rate of 17 Mbps is not adequate to provide perceptually transparent coding of all HDTV inputs. As a result, the system oscillates between the JND and MND modes of Fig. 23, with a constant bit rate constraint in both cases, and no claim is made of overall optimality.

2) Three-Dimensional Subband Coding: Figure 47 shows the result of decomposing a *beach-and-flowers* scene using the 3D subband analysis [70], [152] described in Figs. 12(c) and 14(b). Most of the picture energy is in subband 1, which contains the low-frequency information in the horizontal, vertical, and temporal frequency dimensions. Other subbands have much lower energy. Subband 8 contains high temporal and low spatial frequencies; and therefore acts as a motion detector.

Energy-based adaptive bit allocation provides a classical approach to algorithm optimization. Although straightforward in principle, such optimization is very difficult because of the dynamics of the problem: all energies vary in three dimensions, horizontal, vertical, and temporal, and even if an optimal bit-allocation algorithm is defined, there is the additional problem of transmitting the bit-allocation information to the decoder.

Perceptual optimization of the coder is even more difficult. In recent experimentation with the system for videoconferencing at medium bit rates (such as 384 kbps), empirical perceptual designs have been proposed for the following functions: bit allocation based partly on energy (favoring subband 1) and partly for motion fidelity (favoring subband 8); adaptive exchange of spatial and temporal resolution in subband 1; and perceptually efficient quantization of the higher frequency subbands using geometric vector quantization, as illustrated in Fig. 21(b) [116], [117].

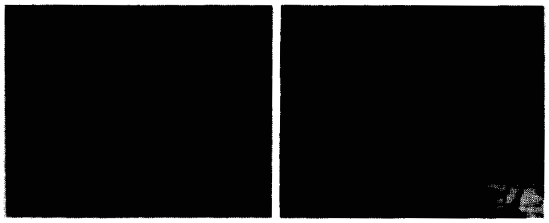


Fig. 47. 3D subband decomposition of *Beach and Flowers* (after [116], [117]). The subband numbers 1 through 11 are labeled in Fig. 14(b).

At the bit rate of 384 kbps, and with inputs in the CIF format (360 × 240 pixels, 15 frames/s), the 3D coder operates well below the transparency level; the JND principle of Fig. 23 does not apply, and the MND cues used in the system are at best informal and empirical. This last observation is also true of MPEG and CCITT video coders operating at similar bit rates, although the nature of the distortion is different in these systems. None of the currently known video algorithms provides high picture quality (a score of 4.0 on the *mos* or impairment scales) at rates much below 1 Mbps. However, useful applications are possible with lower levels of quality, and this has led to a significant degree of commercialization of low bit rate systems for videoconferencing and videotelephony.

X. Research Directions

The technology of signal compression has seen significant advances in recent years, but we still see a gap between current capabilities and technology targets that we feel are attainable (Fig. 5). Several research directions are needed to attain these goals. Perceptual coding is one of them. This concept has made a steady and ever-increasing impact on the status of signal compression. In the future, to enhance the role of perceptual coding and thus that of signal compression in general, we need to address several research challenges, some of which are described below.

Algorithms for perceptual coding need to become more robust, scalable, and portable. One should be able to address different signal types (luminance and chrominance, speech, and audio), different coder types (motion compensation and 3D coding, subband, and transform coding, vector quantization and block fractal coding [61]), and different signal environments (interlaced versus progressive scanning, very-clean versus camera-noise-limited inputs, near- versus far-distance viewing), without having to re-invent a new empirical perceptual model for every situation. If the perceptual algorithms are made more robust in the above

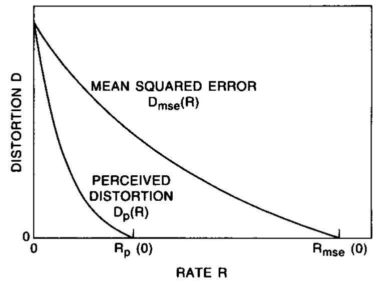


Fig. 48. Distortion-rate function with mean-squared and perceptual error metrics.

ways, there is also a much better chance of incorporating them in rapidly evolving technology for new services and coding standards.

The role of perception in various modules of coding needs to be better understood. In other words, perceptual metrics need to be integrated better into algorithms for motion compensation, spatiotemporal 3D coding, vector quantization, pre- and postprocessing, methods of time–frequency analysis including multiresolution, pyramid, and wavelet techniques, and finally, emerging models for signal production (such as the articulatory model for speech [136] and the wire-frame model for images of the human face [4]).

In audition and vision alike, the theories of noise masking need to be better unified and, in particular, temporal masking needs to be better understood and applied.

The continuum of distortion in the MND model of Fig. 23 needs significant additional work. At this time, it is much easier to establish the JND point (the rate $R_p(0)$ for zero perceptual distortion) than to define the way perceived distortion $D_p(R)$ varies as a function of R in the so-called supra-threshold region of nonzero D. This is shown in Fig. 48. The exact form of $D_p(R)$ is really unknown although we know that it is displaced well to the left of the mse-based distortion-rate curve; this curve has a zero-distortion rate $R_{\rm mse}(0)$ that is significantly higher than the JND point $R_p(0)$. The homomorphic model may prove to be useful in methodologies for determining the shape of the perceptual distortion-rate curve. Even if the $D_p(R)$ function is available, the incorporation of this knowledge in an actual encoder with a nonstationary input is a formidable problem. Part of solving this problem is the design of an efficient buffer control mechanism that provides the bit rate control in Fig. 23.

With increased understanding of perceptual quality metrics (in the internal workings of a coding algorithm), one could also expect better methodologies for evaluating the subjective quality of the final signal (at the output of the decoder). In particular, we can expect more progress in our search for a subjectively meaningful objective measure of overall quality, and we can perhaps minimize the need for time-consuming and intricate subjective tests.

Finally, as perceptual coding evolves into more of a science (rather than the art it currently is), the lessons and the tools from this field will hopefully extend to tasks well beyond signal compression itself. One such problem is that of measuring and maximizing the quality of service in a multiuser network in the presence of various classes of signal degradation: coding distortion, bit error effects, packet losses, and delay. Another challenge is that of measuring and maximizing the quality of composite signals in a multimedia system: the perceived overall quality of an audiovisual signal, as opposed to the individual quality levels of its audio and visual components; and the related bigger problem of evaluating and improving the perceived quality of telepresence in the sophisticated communication systems of the next generation.

Acknowledgment

The authors wish to thank the reviewers of an earlier version of this manuscript, as well as colleagues at AT&T Bell Laboratories for their comments, Dr. J. Schroeter and Dr. J. Kovacevic, in particular. It is also a pleasure to acknowledge several years of inspiring collaboration with a large number of scientists whose outstanding work is reflected in this overview.

References

[1] N. Ahmed, T. Natarajan, and K. Rao, "Discrete cosine transform," *IEEE Trans. Computers*, pp. 90–93, Jan. 1974.

[2] A. J. Ahumada, Jr., "Putting the visual system noise back in the picture," *J. Opt. Soc. Amer. A*, vol. 4, no. 12, pp. 2372–2378, Dec. 1987.

[3] A. J. Ahumada, Jr. and H. A. Peterson, "Luminance-model-based dct quantization for color image compression," in *Proc SPIE Conf. on Human Vision, Visual Processing and Digital Display III*, B. E. Rogowitz, Ed.,1992, pp. 365–374.

[4] K. Aizawa, H. Harashima, and T. Saito, "Model-based analysis-synthesis image coding (mbasic) system for a persons face," in *Signal Processing: Image Communication.* Amsterdam, The Netherlands: Elsevier, Oct. 1989, pp. 139-152.

[5] B. S. Atal, "High-quality speech at low bit rates: multi-pulse and stochastically excited linear predictive coders," *Proc. ICASSP*, pp. 1681-1684, 1986.

[6] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, pp. 247–254, June 1979.

[7] M. Barnsley, *Fractals Everywhere.* New York: Academic Press, 1988.

[8] R. Baseri and V. J. Mathews, "Vector quantization of image using visual masking functions," in *Proc. ICASSP'92*, vol. III (San Francisco, CA, 1992), pp. 365–368.

[9] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, NJ: Prentice Hall, 1971.

[10] R. M. Boynton, *Human Color Vision.* New York: Holt, Rinchart, Winston, 1979.

[11] K. H. Brandenburg, "OCF—A new coding algorithm for high quality sound signals," *Proc. ICASSP*, pp. 141–144, Apr. 1987.

[12] G. Buchsbaum, "Color signal coding: color vision and color television," *Color Res. Appl.*, vol. 12, no. 5, Oct. 1987.

[13] Z. L. Budrikis, "Visual fidelity criterion and modeling," *Proc. IEEE*, vol. 60, no. 7, pp. 771–779, July 1972.

[14] CCIR, "Method for the Subjective Assessment of the Quality of Television Pictures," Rec. 500-1, 1978.

[15] CCIR, "Listening Tests for the Assessment of Low Bit Rate Audio Coding Systems," CCIR Doc. TG 1012-6 (rev. 2), Nov. 22, 1991.

[16] CCITT Study Group XVIII, "7 kHz Audio Coding Within 64 kb/s," CCITT Draft Recommendation G.722, Report of Working Party XVIII/8, July 1986.

[17] F. W. Campbell and J. G. Robson, "Application of Fourier analysis to the visibility of gratings," *J. Physiol.*, pp. 551–566, 1968.

[18] D. Chen and A. Bovik, "Hierarchical visual pattern image coding," submitted to Conf. on Image, Speech and Signal Processing in the 10th Int. Conf. on Pattern Recognition.

[19] J.-H. Chen and A. Gersho, "Real-time vector apc speech coding at 4800 b/s with adaptive postfiltering," *Proc. ICASSP*, pp. 2185–2188, Apr. 1987.

[20] J.-H. Chen, R. V. Cox, Y-C. Lin, N. S. Jayant, and M. J. Melchner, "A Low Delay CELP Coder for the CCITT 16 kbps Speech Coding Standard," June 1992.

[21] W.-H. Chen and W. K. Pratt, "Scene adaptive coder," *IEEE Trans. Commun.*, vol. COM-32, no. 3, pp. 225–232, Mar. 1984.

[22] B. Chitprasert and K. R. Rao, "Human visual weighted progressive image transmission," *IEEE Trans. Commun.*, vol. 38, pp. 1040–1044, July 1990.

[23] T. N. Cornsweet, *Visual Perception* New York: Academic Press, 1970.

[24] D. C. Cox, "Portable digital radio communications—An approach to tetherless assess," *IEEE Commun. Mag.*, pp. 30–40, July 1989.

[25] R. V. Cox, "The design of uniformly and nonuniformly spaced pseudoquadrature mirror filters," *IEEE Trans. Acoust., Speech, Signal Process.*, pp. 1090-1096, 1986.

[26] R. V. Cox, J. Hagenauer, N. Seshadri, and C-E. W. Sundberg, "Subband speech coding and matched convolutional channel coding for mobile radio channels," *IEEE Trans. Signal Process.*, vol. 39, no. 8, pp. 1717–1731, Aug. 1991.

[27] R. E. Crochiere, S. A. Webber, and J. L. Flanagan, "Digital coding of speech in subbands," *Bell Syst. Tech. J.*, pp. 1069–1085, 1976.

[28] C. C. Cutler, "Differential Quantization for Communication Signals," U.S. Patent 2 605 361, July 29, 1952.

[29] —— "Delayed encoding: stabilizer for adaptive coders," *IEEE Trans. Commun.*, pp. 898–904, Dec.1971.

[30] S. Daly, "The visual difference predictor: an algorithm for the assessment of image fidelity," in *Proc. SPIE Conf. on Human Vision, Visual Processing and Digital Display III* (San Jose, CA, 1992), SPIE #1666, pp. 2–15.

[31] I. Daubechies, "Orthonormal bases on compactly supported wavelets," *Commun. Pure Appl. Math.*, pp. 909–996, 1988.

[32] W. R. Daumer, "Subjective evaluation of several efficient speech coders," *IEEE Trans. Commun.*, pp. 655–662, Apr. 1982.

[33] L. D. Davisson, "Rate distortion theory and application," *Proc. IEEE*, vol. 70, pp. 800–808, July 1972.

[34] D. Esteban and C. Galand, "Application of quadrature mirror filters to split band voice coding schemes," *Proc. ICASSP*, pp. 191–195, 1987.

[35] O. Faugeras, "Digital color image processing within the framework of a human visual model," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, Aug. 1979.

[36] J. L. Flanagan, *Speech Analysis, Synthesis and Perception.* New York: Springer-Verlag, 1972.

[37] J. L. Flanagan, M. R. Schroeder, B. S. Atal, R. E. Crochiere, N. S. Jayant, and J. M. Tribolet, "Speech coding," *IEEE Trans. Commun.*, pp. 710–737, Apr. 1979.

[38] N. Fletcher, "Auditory patterns," *Rev. Mod. Phys.*, pp. 47–65, 1940.

[39] R. Forscheimer and T. Kronander, "Image coding — from waveforms to animation," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 37, pp. 2008–2023, Dec. 1989.

[40] A. Fuldseth, E. Harborg, F. T. Johansen, and J. E. Knudsen, "A real time implementable 7 khz speech coder at 16 kbps," in *Proc. Eurospeech 91*(Genoa, Italy, 1991).

[41] R. G. Gallager, *Information Theory and Reliable Communication.* New York: McGraw Hill, 1965.

[42] A. Gersho, "Asymptotically optimum block quantization," *IEEE Trans. Informat. Theory*, pp. 373–380, July 1979.

[43] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Dordrecht, The Netherlands: Kluwer, 1992.

[44] I. A. Gerson and M. A. Jasiuk, "Vector sum excited linear prediction (vselp)," presented at IEEE Workshop on Speech Coding for Telecommunication, Sept. 5-8, 1989.

[45] C. Gerwin and T. Ryden, "Subjective assessments on low bitrate audio codecs," in *Proc. 10th Internat. AES Conf. on Images of Audio* (London, England, Sept. 1991), pp. 91–102.

[46] J. D. Gibson, "Sequentially adaptive backward prediction in adpcm speech coders," *IEEE Trans. Commun.*, pp. 145–150, Jan. 1978.

[47] B. Girod, "The information theoretical significance of spatial and temporal masking in video signals," in *Proc. SPSE/SPIE Symp. on Human Vision, Visual Processing and Display* (Los Angeles, CA, Jan. 1989).

[48] R. C. Gonzalez and P. Wintz, *Digital Image Processing.* Reading, MA: Addison-Wesley, 1977.

[49] D. J. Goodman, "Embedded DPCM for variable bit rate transmission," *IEEE Trans. Commun.*, pp. 1040–1066, July 1980.

[50] ——, "Speech quality of the same speech transmission conditions in seven different countries," *IEEE Trans. Commun.*, pp. 642–654, Apr. 1982.

[51] D. J. Granrath, "The role of human visual models in image processing," *Proc. IEEE*, vol. 69, pp. 552–561, May 1981.

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

[53] R. M. Gray, P. C. Cosman, and K. L. Oehler, "Incorporating visual factors into vector quantizers for image compression," to appear in *Visual Factors in Electronic Image Communications*, A.B. Watson, Ed. Boston, MA: MIT Press.

[54] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust. Speech, Ssignal Process.*, vol. 36, pp. 1223–1235, 1988.

[55] C. F. Hall and E. L. Hall, "A nonlinear model for the spatial characteristics of the human visual system," *IEEE Trans. Syst. Man. Cybern.*, pp. 161–170, Mar. 1977.

[56] R. P. Hellman, "Asymmetry of masking between noise and tone," *Perception and Psychophysics*, vol. 11, pp. 241–246, 1972.

[57] J. J. Y. Huang and P. M. Schultheiss, "Block quantization of correlated gaussian random variables," *IRE Trans. Commun. Syst.*, vol. CS-11, pp. 289–296, 1963.

[58] D. Huffman, "A method for the construction of minimum redundancy codes," *Proc. IRE*, vol. 40, pp. 1098–1101, Sept. 1952.

[59] ISO: "Coded Representation of Picture and Audio Information—Progressive Bi-Level Image Compression Standard," ISO/IEC draft, Dec. 14, 1990.

[60] ISO/IEC JTCI/SC29, "Information Technology—Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbps — CD 11172 (Part 3, Audio)," Doc. ISO/IEC JTCI/SC29 NO71, Dec. 1991.

[61] A. E. Jacquin, "A novel fractal block-coding technique for digital images," *Proc. ICASSP*, pp. 2225–2228, 1990.

[62] N. S. Jayant, "Signal compression: technology targets and research directions," *IEEE J. Selected Areas Commun.* (Special Issue on Speech and Image Coding), June 1992.

[63] ——, *Waveform Quantization and Coding* New York: IEEE Press, 1976.

[64] N. S. Jayant, J. D. Johnston, and Y. Shoham, "Coding of wideband speech," in *Proc. 2nd European Conf. on Speech Communication and Technology*, Sept. 1991.

[65] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video.* Englewood Cliffs, NJ: Prentice Hall, 1984.

[66] J. D. Johnston, "A filter family designed for use in quadrature mirror filter banks," *Proc. ICASSP*, 1980.

[67] ——, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Selected Areas Commun.*, pp. 314–323, Feb. 1988.

[68] J. D. 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.

[69] J. D. Johnston and A. Ferreira, "Sum-difference stereo transform coding," *Proc. ICASSP*, pp. II-569–II-572, Apr. 1992.

[70] G. Karlsson and M. Vetterli, "Three dimensional subband coding of video," *Proc. ICASSP*, 1988.

[71] G. Karlsson and M. Vetterli, "Packet video and its integration into the network architecture," *IEEE J. Selected Areas Commun.*, pp. 739–751, June 1989.

[72] D. P. Kemp, R. A. Sueda, and T. E. Tremain, "An evaluation of 4800 b/s voice coders," *Proc. ICASSP*, May 1989.

[73] T. Kim, "New finite state vector quantizers for images," *Proc. ICASSP*, pp. 1180–1183, 1988. Also, *IEEE Trans. Image Process.*, 1992.

[74] N. Kitawaki, M. Honda, and K. Itoh, "Speech-quality assessment methods for speech-coding systems," *IEEE Commun. Mag.*, pp. 26–32, Oct. 1984.

[75] N. Kitawaki and K. Itoh, "Pure delay effects on speech quality in telecommunications," *IEEE J. Selected Areas Commun.*, pp. 586–593, May 1991.

[76] N. Kitawaki and H. Nagabuchi, "Quality assessment of speech coding and speech synthesis systems," *IEEE Commun. Mag.*, pp. 36–44, Oct. 1988.

[77] N. Kitawaki, H. Nagabuchi, M. Taka, and K. Takahashi, "Speech coding technology for atm networks," *IEEE Commun. Mag..*, pp. 21–27, Jan. 1990.

[78] S. A. Klein, A. D. Silverstein, and T. Carney, "Relevance of human vision to jpeg-dct compression," in *Proc. SPIE Conf. on Human Vision, Visual Processing and Digital Display III* (San Jose, CA, 1992), SPIE #1666, pp. 200–215.

[79] E. T. Klemmer, "Subjective evaluation of transmission delay in telephone conversations," *Bell Syst. Tech. J.*, pp. 1141–1146, July–Aug. 1967.

[80] F. J. Kolb, Jr., Ed. "Bibliography: Psychophysics of image evaluation," *SMPTE J.*, pp. 594–599, Aug. 1989.

[81] S. Komiyama, "Subjective evaluation of angular displacement between picture and sound directions for HDTV sound systems," *J. Audio Eng. Soc.*, vol. 37, no. 4, pp. 210–214, Apr. 1989.

[82] J. Kovacevic, "Subband coding systems incorporating quantizer models," in *Proc. Data Commun. Conf.* (Snowbird, Mar. 1993) Also submitted to *IEEE Trans. Image Process.*.

[83] P. Kroon, E. F. Deprettere, and R. J. Sluyter, "Regular-pulse excitation — A novel approach to effective and efficient multipulse coding of speech," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-34, no. 5, pp. 1054–1063, Oct. 1986.

[84] M. Kunt, A. Ikonmopoulos, and M. Kocher, "Second-generation image coding technique," *Proc. IEEE*, vol. 73, pp. 549–574, Apr. 1985.

[85] D. LeGall, "MPEG, A video compression standard for multimedia applications," *Commun. ACM*, pp. 47-58, Apr. 1991.

[86] J. O. Limb, "Distortion criteria of the human viewer," *IEEE Trans. Systems, Man, Cybern.*, pp. 788–793, Dec. 1979.

[87] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantization design," *IEEE Trans. Commun.*, pp. 84–95, Jan. 1980.

[88] M. Liou, "Overview of the p*64 kbits/s video coding standard,"*Commun. ACM*, pp. 60-63, Apr. 1991.

[89] S. S. Magan, "Trends in DSP system design," in short course on *Digital Signal Processing* (IEEE Int. Electron Device Meet., Dec. 1989).

[90] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," *Proc. IEEE*, vol. 73, pp. 1551–1588, Nov. 1985.

[91] H. S. Malvar, *Signal Processing with Lapped Transforms.* Dedham, MA: Artech House, 1992.

[92] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Informat. Theory*, vol. IT-20, no. 4, July 1974.

[93] J. Max, "Quantizing for minimum distortion," *IRE Trans. Informat. Theory*, vol. IT-6, pp. 7–12, Mar. 1960.

[94] R. J. McAulay and T. F. Quatieri, "Speech analysis and synthesis based on a sinusoidal model," *IEEE Trans. Acoust. Speech Ssignal Process.*, pp. 744–754, Aug. 1986.

[95] D. L. McLaren and D. T. Nguyen, "Removal of subjective redundancy from dct-coded images," *Proc. Inst. Elec. Eng.*, pt. I, pp. 345–350, Oct. 1991.

[96] P. Mermelstein, "G.722, A new CCITT coding standard for digital transmission of wideband audio signals," *IEEE Commun. Mag.*, pp. 8–15, Jan. 1988.

[97] G. Morrison and D. Beaument, "Two-level video coding for ATM networks," *Signal Processing: Image Communication*, pp. 179–195, June 1991.

[98] H. G. Musmann, "The ISO audio coding standard," in *Proc. IEEE Globecom Conf.*, Dec. 1990.

[99] H. G. Musmann, P. Pirsch, and H.-J. Grallert, "Advances in picture coding," *Proc. IEEE*, vol. 73, pp. 523–548, Apr. 1985.

[100] A. N. Netravali and B. G. Haskell, *Digital Pictures: Representation and Compression.* New York: Plenum, 1988.

[101] A. N. Netravali and J. O. Limb, "Picture coding: A review," *Proc. IEEE*, vol. 68, pp. 366-406, Mar. 1980.

[102] A. N. Netravali, E. Petajan, S. Knauer, K. Mathews, R. J. Safranek, and P. Westerink, "A high quality digital HDTV codec," *IEEE Trans. Consumer Electron.*, pp. 320–330, Aug. 1991.

[103] A. N. Netravali and J. A. Stuller, "Motion compensation transform coding," *Bell Syst. Tech. J.*, pp. 1703–1718, Sept. 1974.

[104] D. L. Neuhoff and T. N. Pappas, "Perceptual coding of images for halftone display," *Proc. ICASSP*, May 1991.

[105] K. N. Ngan, K. S. Leong, and H. Singh, "Adaptive cosine transform coding of images in perceptual domain," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 37, no. 11, pp. 1743–1750, Nov. 1989.

[106] N. B. Nill, "A visual model weighted cosine transform for image compression and quality assessment," *IEEE Trans. Commun.*, p. 551, June 1985.

[107] Y. Ninomiya, "HDTV broadcasting systems," *IEEE Commun. Mag.*, pp. 15–23, Aug. 1991.

[108] P. Noll, "High quality speech and wideband audio coding," *IEEE Commun. Mag.*, Nov. 1993.

[109] M. Nomura, T. Fujii, and N. Ohta, "Layered coding for ATM based video distribution systems," *Signal Processing: Image Communication*, pp. 301–311, 1991.

[110] J. B. O'Neal, "Predictive quantizing systems for the transmission of television signals," *Bell Syst. Tech. J.*, pp. 689–719, May–June 1966.

[111] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proc. IEEE*, vol. 69, pp. 529–541, May 1981.

[112] T. N. Pappas, "Perceptual coding and printing of gray-scale and color images," in *SID-92 Dig. Tech. Papers* (Boston, MA, May 1992).

[113] W. B. Pennebaker and J. L. Mitchell, *JPEG Still Image Data Compression Standard.* New York: Van Nostrand Reinhold, 1992.

[114] M. G. Perkins and T. Lookabaugh, "A psychophysically justified bit allocation algorithm for subband image coding systems," *Proc. ICASSP*, 1989.

[115] H. A. Peterson, "DCT basis function visibility in rgb space," in *Soc. Informat. Display Dig. Tech. Papers*, J. Moreale, Ed. Playa del Ray, CA: Soc. Informat. Display.

[116] C. I. Podilchuk and N. Farvardin, "Perceptually based low bit rate video coding," *Proc. ICASSP*, 1991.

[117] C. I. Podilchuk, N. S. Jayant, and P. Noll, "Sparse codebooks for the quantization of non-dominant sub-bands in image coding," *Proc. ICASSP*, 1990.

[118] W. K. Pratt, *Digital Image Processing.* New York: Wiley, 1978.

[119] J. Princen, A. Johnson, and A. Bradley, "Sub-band transform coding using filterbank designs based on time-domain aliasing cancellation," *Proc. ICASSP*, pp. 2161–2164, 1987.

[120] A. Puri and R. Aravind, "Motion-compensated video coding with adaptive perceptual quantization," *IEEE Trans. Circuits Syst.Video Technol.*, vol. 1, no. 4, Dec. 1991.

[121] S. R. Quackenbush, "Hardware implementation of a color image decoder for remote database access," *Proc. ICASSP*, 1990.

[122] L. R. Rabiner and R. W. Schafer, *Digital Speech Processing.* Englewood Cliffs, NJ: Prentice Hall, 1978.

[123] V. Ramamoorthy, N. S. Jayant, R. V. Cox, and M. M. Sondhi, "Enhancement of ADPCM speech coding with backward-adaptive algorithms for postfiltering and noise feedback," *IEEE J. Selected Areas in Commun.*, pp. 364–382, Feb. 1988.

[124] R. R. Riesz and E. T. Klemmer, "Subjective evaluation of delay and echo suppressors in telephone communications," *Bell Syst. Tech. J.*, pp. 2919–2943, 1963.

[125] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Process. Mag.*, pp. 14–38, Oct. 1991.

[126] L. G. Roberts, "Picture coding using pseudo-random noise," *IRE Trans. Informat. Theory*, pp. 145–154, Feb. 1962.

[127] G. Roy and P. Kabal, "Wideband CELP speech coding at 16 kbps," *Proc. ICASSP*, pp. 17–20, 1991.

[128] R. J. Safranek and J. D. Johnston, "A perceptually tuned sub-band image coder with image-dependent quantization and post-quantization data compression," *Proc. ICASSP*, 1989.

[129] R. J. Safranek, J. D. Johnston, and R. E. Rosenholtz, "A perceptually tuned sub-band image coder," presented at the SPIE Symp. on Electronic Imaging Human Vision and Electronic Imaging: Models Methods and Applications, Santa Clara, CA, Feb. 12–14, 1990.

[130] J. A. Saghri, P. S. Cheatham, and A. Habibi, "Image quality measure based on a human visual system model," *Opt. Eng.*, vol. 28, no. 7, pp. 813–818, July 1989.

[131] D. J. Sakrison, "Image coding applications of vision models," in *Image Transmission Techniques*, W. K. Pratt, Ed. New York: Academic Press, May 1979, pp. 21-51.

[132] B. Scharf, "Critical bands," in *Foundations of Modern Auditory Theory*, J. Tobias, Ed. New York: Academic Press, 1970, pp. 159–202.

[133] W. F. Schreiber, "Psychophysics and the improvement of television picture quality," *SMPTE J.*, pp. 717-725, Aug. 1984.
[134] W. F. Schreiber and A. B. Lippman, "Reliable EDTV/HDTV transmission in low-quality analog channels," ATRP-T-96R, SMPTE Talk, Oct. 16, 1988.
[135] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, vol. 66, no. 6, pp. 1647–1651, Dec. 1979.
[136] J. Schroeter and M. M. Sondhi, Eds. "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Ed. New York: Dekker, 1991.
[137] C. E. Shannon, "A mathematical theory of communications," *Bell Syst. Tech. J.*, vol. 27, pp. 379-423, 623–656, 1948.
[138] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *IRE Nat. Conv. Rec.*, pt. 4, pp. 142–163, 1959.
[139] Y. Shoham, "Constrained excitation coding of speech at 4.8 kbps," in *Advances in Speech Coding*. Dordrecht, The Netherlands: Kluwer, 1991, pp. 339–348.
[140] B. Smith, "Instantaneous companding of quantized signals," *Bell Syst. Tech. J.*, pp. 653–709, May 1957.
[141] R. Steele, *Delta Modulation Systems*. New York: Halsted, 1975.
[142] ——, "The cellular environment of lightweight handheld portables," *IEEE Commun. Mag.*, pp. 20–29, July 1989.
[143] T. G. Stockham, "Image processing in the context of a visual model," *Proc. IEEE*, vol. 60, no. 7, pp. 828–842, July 1972.
[144] C. F. Stromeyer and B. Julesz, "Spatial frequency masking in vision: critical bands and spread of masking," *J. Opt. Soc. Amer.*, pp. 1221–1232, Oct. 1972.
[145] Swedish Radio, Unpublished report on the quality of low rate audio algorithms submitted for the ISO-MPEG Std., Aug. 1990.
[146] G. Theile, G. Stoll, and M. Link, "Low bit-rate coding of high-quality audio signals," *EBU Tech. Rev.*, no. 230, pp. 71–94, Aug. 1988.
[147] L. Thibault and T. Grusec, "CCIR Listening Test Report: Subjective Evaluation of the Basic Audio Quality of Contribution and Distribution Codecs," CRC Rep., Commun. Res.Ctr., Ottawa,Ont., Canada, 1992.
[148] T. E. Tremain, "The Government standard linear predictive coding algorithm: LPC-10," *Speech Technol.*, vol. 1, no. 2, pp. 40–49, Apr. 1982.
[149] P. P. Vaidyanathan, "Quadrature mirror filter banks, m-band extensions and perfect-reconstruction techniques," *IEEE ASSP Mag.*, pp. 4–20, 1987.
[150] R. E. Van Dyck and S. A. Rajala, "Subband/VQ coding in perceptually uniform color spaces," *Proc. ICASSP*, pp. III-237–III-240, 1992.
[151] W. Verbiest and L. Pinnoo, "A variable bit rate video coder for asynchronous transfer mode networks," *IEEE J. Select. Areas Commun.*, pp. 761–770, June 1989.
[152] M. Vetterli, "Multidimensional sub-band coding: some theory and algorithms," *Signal Process.*, pp. 97–112, Apr. 1984.
[153] M. Vetterli and C. Herley, "Wavelets and filter banks: theory and design," *IEEE Trans. Signal Process.*, Sept. 1992.
[154] W. D. Voiers, "Diagnostic Acceptability measure for speech communication systems," *Proc. ICASSP*, pp. 204–207, May 1977.
[155] W. D. Voiers, "Diagnostic evaluation of speech intelligibility," in *Speech Intelligibility and Speaker Recognition*, M. Hawley, Ed. Stroudsburg, PA: Dowden Hutchinson Ross, 1977.
[156] G. K. Wallace, "The JPEG still picture compression standard,"*Commun. ACM*, pp. 31–43, Apr. 1991.
[157] L. Wang and M. Goldberg, "Progressive image transmission using vector quantization on images in pyramid form," *IEEE Trans. Commun.*, pp. 1339–1349, Dec. 1989.
[158] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Select. Areas Commun.*, pp. 819–829, June 1992.
[159] Y. Wang and Q-F. Zhu, "Signal loss recovery in dct-based image and video codecs," *Proc. SPIE Visual Commun. Image Process.*, pp. 667–678, 1991.
[160] H. Watanabe, "Integrated office systems: 1995 and beyond," *IEEE Commun. Mag.*, pp. 74–80, Dec. 1987.
[161] A. B. Watson, "The Cortex Transform: Rapid computation of simulated neural images," *Comput. Vision, Graphics, Image Process.*, vol. 39, no. 3, pp. 311–327, 1987.
[162] A. B. Watson, "Efficiency of an image code based on human vision," *J. Opt. Soc. Amer. A*, vol. 4, no. 12, pp. 2401–2417, 1987.
[163] A. B. Watson, "Visually optimal DCT quantization matrices for individual images," in J. A. Storer and M. Cohn, Eds., *Proc. Data Compression Conf.*, IEEE Computer Society Press, 1993, pp. 178–187.
[164] S. A. Webber, C. J. Harris and J. L. Flanagan, "Use of Variable quality coding and time interval modification in packet transmission of speech," *Bell Syst. Tech. J.*, vol. 56, pp. 1569–1573, Oct. 1977.
[165] J. H. D. M. Westerink, "Subjective image quality as a function of viewing distance, resolution and picture size," IPO Annual Progress Rep. 22, pp. 55–64, 1987.
[166] C. Wierzynski, "Improved time domain noise shaping based on perceptual criteria," M.S. thesis, MIT, June 1992.
[167] R. Wilson, H. E. Knutsson, and G. H. Granlund, "Anisotropic nonstationary image estimation and its applications: Part II—Predictive image coding," *IEEE Trans. Commun.*, pp. 398–406, Mar. 1983.
[168] P. A. Wintz, "Transform picture coding," *Proc. IEEE*, vol.60, pp. 809–820, July 1972.
[169] I. H. Witten, R. M. Neal, and J. G. Cleary, "Arithmetic Coding for data compression," *Commun. ACM*, pp. 520–540, June 1987.
[170] S. Wolf, C. A. Dvorak, R. F. Kubichek, C. R. South, R. A. Schaphorst, and S. D. Voran, "How will we rate telecommunications system performance?", *IEEE Commun. Mag.*, pp. 23–29, Oct. 1991.
[171] J. W. Woods and S. D. O'Neil, "Subband coding of images," *IEEE Trans. Acoust., Speech Signal Process.*, pp. 1278–1288, Oct. 1986.
[172] A. D. Wyner, "Fundamental limits in information theory," *Proc. IEEE*, vol. 69, pp. 239–251, Feb. 1981.
[173] R. Zelinski and P. Noll, "Adaptive transform coding of speech signals," *IEEE Trans. Acoust., Speech, Signal Process.*, pp. 299–309, Aug. 1977.
[174] E. Zwicker, *Psychoakustik* (in German). New York: Springer-Verlag, 1982.
[175] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. New York: Springer-Verlag, 1990.

Nikil Jayant (Fellow, IEEE) received the Ph.D. degree in electrical communications engineering from the Indian Institute of Science, Bangalore, India.

He joined AT&T Bell Laboratories in 1968 and is now Head of Signal Processing Research Department at Murray Hill, NJ. He is responsible for research in speech and image processing with application to coding, communications, and recognition. Current research programs in his department include signal coding with human perceptual models; techniques for source coding, channel coding, and modulation; automatic document recognition; and real-time implementation of signal processing algorithms in support of technology for digital cellular radio, advanced television, personal communications services, and Integrated Services Digital Networks. He is also Head of the Advanced Audio Technology Department at AT&T. In this position, he is responsible for developing technology for audio compression and digital audio broadcasting. He is the Editor of *Waveform Quantization and Coding* (New York: IEEE Press, 1976) and co-author of *Digital Coding of Waveforms—Principles and Applications to Speech and Video* (Englewood Cliffs, NJ:Prentice-Hall, 1984)

Dr. Jayant was the first Editor-in-Chief of the IEEE *ASSP Magazine*, a publication of the Acoustics, Speech, and Signal Processing Society of the IEEE and received the IEEE Browder J. Thompson Memorial Prize Paper Award.

[redacted]

James D. Johnston received the B.S.E.E. and M.S.E.E. degrees from Carnegie-Mellon University, Pittsburgh, PA, in 1975 and 1976, respectively.

Since 1976, he has been with AT&T Bell Laboratories in the Acoustics Research and Signal Processing Research Departments. He has worked in analog signal processing, speech coding, quadrature mirror filter design, and perceptual coding of both audio and image. His current interests are multichannel coding and soundfield reconstruction of audio signals, and a comprehensive visual model for still-frame images.

[redacted]

Robert J. Safranek (Senior Member, IEEE) was born in Manitowoc, WI, in 1958. He received the B.S.E.E. degree in 1980, the M.S.E.E. degree in 1982, and the Ph.D. degree in 1986, all from Purdue University, West Lafayette, IN.

Since 1986, he has been a member of the Signal Processing Research Department at AT&T Bell Laboratories, Murray Hill, NJ, where he has worked on a variety of problems in the areas of perceptual coding, visual modeling, digital video, and HDTV. His research interests include human and machine perception, video-processing techniques, multimedia, and hardware/software systems for signal processing

Dr. Safranek is a member of Eta Kappa Nu and SMPTE.