クロマ：プロ映像／音響の最先端技術誌

1993年1月1日発行（毎月1回1日発行）第8巻第1号通巻64号
平成元年12月8日第3種郵便物認可

CHROMA

特集＝'92国際放送機器展

●ビット圧縮＝2——プロオーディオ最前線
●CCDスタジオカメラ、ソニーBVP-370、BVP-70ISについて
●HDレポート：『NHKヒットステージ』のサイマル放送について
●松下電器産業/情報通信システムセンター
●CG REPORTS：NICOGRAPH'92
●照明の実際とウラ知識：鏡の中のライティング

1 1993

デジタルオーディオと低ビットレート・コーディング

ドルビー研究所東京連絡所
伏木 雅昭

デジタル技術には最近確実に新しい大きなうねりが起こっているが､その総称としての『ビット圧縮』という表現は多分に心理的偏見を助長する響きを含んでいる。現実にこうした新技術はアナログ的圧縮・伸張とは次元を異にするコーディング手法であり、オーディオを決して疎かにしたくない人々にもむしろ冷静にそのコーディング技術を理解してもらい、且つ我々としてはこの技術の持つ積極的意義を重視して『低ビットレート・コーディング』とここでは呼ぶことにしたい。この技術は例えば、機能に違いはあるがパソコン・アーカイバのように、生活必需品としての将来性が期待される。低ビットレート=高効率伝送は特定の信号対象に対するデジタル処理の完成度を計る上でも欠かすことの出来ない技術要素である。

写真1　DSTL5501エンコーダー(上)とDSTL5502デコーダー

開発目標

低ビットレート・コーディングは最新の聴覚器宮モデルとそれに関連する耳の可聴限界とを応用した処理技術で、ドルビーAC-2 を始め現在最良とされる低ビットレート・コーダーではチャンネル当り 128kB/s のデータレートでコンパクトディスクに匹敵する音質を提供できる。従って、その効率は CD の約 5 倍となる。これを支えているのが DSP やカスタム LSI などの集積素子で、業務用・民生用を問わず通信・放送・高品位テレビ・光学及び磁気録音さらにはコンピュータマルチメディア製品などの用途にコーダーを搭載するのに必要とされる演算能力と経済性とが実現可能となったことから、技術の具現化に大きく拍車がかかった。

低ビットレート・オーディオ・コーダー開発の第一目標はコスト効率の高い高品質デジタル音声の伝送・記録を行なうことにあり、会話用などのコーディング技術が人間の声という限定された質の音を対象に明瞭度を何よりも重視しているのとは対照的に、低ビットレート・オーディオ・コーダーの場合には限りなく多彩な自然音や合成音に対して忠実度が損なわれないよう動作できる設計をしなければならない。『認識』の共通項は人間の耳であり、低ビットレート・コーダーでは人間の聴覚反応の可聴限を探ることでデータレートの低減を行なう。

マスキングの概念

『臨界帯域』Critical band の概念[H.フレッチャー,1940 年]及び音響心理的『聴感マスキング』の原理は高性能低ビットレート・オーディオ・コーダーを設計する基本である。聴感マスキングとは、大きな音がその音に帯域的に近接する他の小さな音があるのを聴こえなくする、つまり『マスク』する傾向を持つ現象を言う。マスキングは大信号の周波数帯域で耳の認識スレッシュホールドが上昇して、同居する弱音信号成分のレベルを超えてしまうことに起因している。低レベル信号での実験では、弱い音はそれをマスクする大きな音との周波数差が一定のスレッシュホールド幅を超えるまでは聴こえないという結果が出ている。スレッシュホールド幅、言い替えれば周波数単位での『マスキング選択度』の違いを測定し、それを概念化したのがドルビーAC-2 でも採用している挟帯域幅フィルター並列バンクによる臨界帯域モデルである。このモデルでは臨界帯域フィルターは 500Hz 以下では約 100Hz、500Hz より上では中心周波数の 1/5 という一定の分数比幅になっている。このモデルがサブバンド方式のような周波数リニアな固定幅を持たないのは人間の聴感構造自体が周波数に対数的に反応することを思い出してもらえばよいだろう。つまり、オーディオ・コーダーの設計者にとって重要な原理は耳が 30 バンド・リアルタイム・スペクトラムアナライザーのような機能を持ち、20-20kHz の帯域の中で各バンドの帯域幅も感度スレッシュホールドも均一ではないことだ。単音マスキングの実験では可聴スレッシュホールド(つまり耳が感じることの出来る最も弱い音)より 30dB 上までの範囲では、マスキング効果は最小範囲に留まっていることが解る。しかし、それを超えて音が大きくなるほど、マスキングは広い周波数帯域に及び、従ってその効果は複数の臨界帯域に亙る。マスキング信号より上の周波数帯域に対しては特にそれが顕著となる。

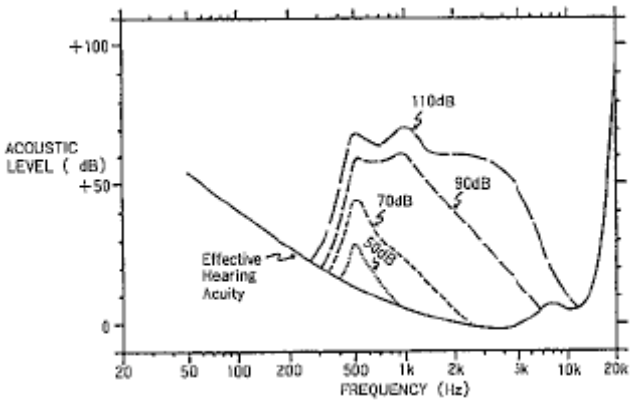

図1　マスキング・スレッシュホールド

低ビットレート・コーダーを最適化する上でさらに考慮すべきマスキングの形態としては過渡信号によるものがある。定常信号状態では、人間の耳は優れた周波数分解能を持つ。しかし、耳が信号変化に同調するにはある程度の時間幅が必要となる。つまり時間分解能では本質的限界があるということである。高レベル過渡テスト信号のオンセット直前の時間幅の実測では、時間分解能は 5 ミリ秒弱である。興味深い点は過渡信号発生の前、最中、後のそれぞれに弱信号に対するマスキングが得られるということだ。『時前マスキング』つまり過渡直前の作用は過渡信号の前 10mS まで有効ある。マスキングは当然ながら過渡信号下が最:も強く、その後 50-200mS の時間内で効果が失われて行く。

オーディオ信号はほぼ一定の持続的信号成分と時間経過とともに急激に変化する過渡成分とで構成されている。時間上ゆっくりと変化する信号は周波数選択度の高いフィルターバンクを用いて、コーディングエラーが信号帯域内に限定され、マスキングが最も効果を持つようコーディングされるのが最善である。しかし過渡信号は耳の時間マスキング範囲を超えてコーディングエラーが時間的に拡散されることのないよう、耳と同じ時間分解能のフィルターバンクを使用してコーディングしなければならない。
優れた周波数分解能と精細な時間分解能とは互いに排他的条件を持つので、フィルターバンクの設計は時間及び周波数マスキングの制約に見合うだけの充分なビット割当を行ない時間及び周波数分解能の間で妥協点を見出すか、処理する信号の特性に応じてその都度時間あるいは周波数分解能のいずれかに最適化できる経時可変型を取るかの方法を用いなければならない。

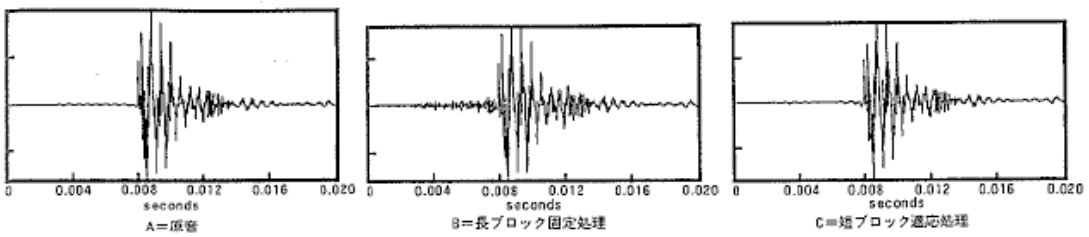

図 2　カスタネット信号の時間分析例
(A=原音/B=長ブロック固定処理/C=短ブロック適応処理)

コーディング・プロセス

さて、低ビットレート・コーディングの基本的プロセスだが、これには時間領域でデジタル化されたオーディオ信号の周波数領域データへの転換、聴感マスキングモデルに基づいた粗い精度での信号周波数成分の可変量子化、量子化器に対する様々な要求に適合させたビット割当、そしてコーディングされたデータを伝送もしくは保存した後に行なわれるオリジナルの時間領域波形への近似的再合成である。信号の周波数領域データ化は多数の周波数帯域フィルターバンクを用いることで行なわれる。その際、周波数分解の手法としてはディスクリート・フーリエ変換を基本とするものか、ポリフェーズ・デジタルフィルターを基本とするもののいずれかがもっぱら採用されている。フィルターバンク、周波数分析、マスキング・スレッシュホールドの演算、量子化及びビット割当機能などは DSP や

LSI 技術によって構成されている。

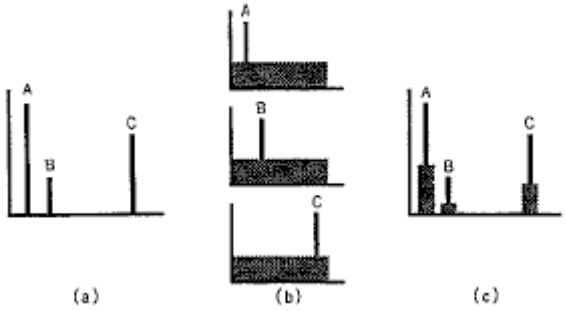

図 3　低データレート・コーダー概念図

マルチバンドの低データレート・コーダー概念図を図 3 に示す。ここでは A、B、C の周波数成分を持つ時間領域信号が標本化・量子化され、臨界帯域周波数選択度を持つ妥当なフィルターバンク技術を用いて周波数領域に変換される。図 3A が変換された信号の周波数領域データである。信号の周波数成分が特定されると、バンド毎に演算:もしくは聴感モデル・プログラムとの比較によって、マスキング・スレッシュホールド概算が得られる。各バンドの最大信号成分を基にマスキング・スレッシュホールドを割り出したら、次に同じ臨界帯域内の弱音成分に対して最大信号のマスキングが充分な効果を持つかどうかの分析が行なわれる。この判定結果が得られれば、後はビットレート低減のプロセスを実行することになる。そのプロセスとは、それぞれのフィルター帯域の成分を量子化雑音が演算された帯域内マスキング・スレッシュホールド値を丁度わずかに下回るだけの精度で量子化し、それら成分のピーク値を補正出来るよう信号の増幅度スケーリングを行なって、デジタル信号処理回路のダイナミックレンジを最適利用することである。注意すべきは、信号の低ビットレート形式データを生成するときに、周波数領域成分の量子化過程で広帯域雑音成分が発生してしまうことだ(図 3B 参照)。先に論じたことでお解りと思うが、この広帯域雑音成分は信号成分によってはマスキングしきれない。そこで、信号の符号化データを伝送もしくは記録後にエンコーダで用いたものと同じフィルターバンクをデコード過程にも使用すれば､量子化された周波数領域データがデコーダのフィルターバンクを通過し、信号の周波数成分はそのままで、しかも量子化雑音成分はマスキング・スレッシュホールド以下の狭い帯域だけに押え込むことが出来る(図 3C 参照)。ここで強調しておきたいことは、低ビットレート・コーダーが行なっている処理の本質は量子化エラーを可聴限下に追いやることであって、可聴限下成分を伝送しないことで量子化ビットが小さいことの問題を回避しているわけである。

AC-2 の製品化

ドルビー研究所では伝送レートによってコーディングの用途及びグレードを一応 3 種類に大別している。つまり、384k(2x192k)=Contribution、256k(2x128k)=Transmission、192k(2x96k)=Emission という定義で、それぞれ「原音グレードの厳格な用途」、「標準的な業務用伝送目的」、「広範なサービス用」とでも受け取ってもらえればよい。現在商品化されているのはこのうち 256k のデータレートに基づくもので、様々な製品が登場し始め

ているので、ここで簡単に紹介しておこう。AC-2 コーディング自体のカスタマイズ・ライセンスも行なっている。

DP-501/502

AC-2 コーディングのステレオ 2 チャンネル汎用システムとして販売された最初の製品である。それぞれ 1U ラックサイズで、DSP エンジンにはモトローラ 56001 を 27MHz で使用している。10 月初旬サンフランシスコで開催された AES コンベンションでは ISDN 回線を使った伝送デモも行なわれていた。

DSTL5501/5502

デジタル伝送の新たな用途のひとつとしてドルビー研究所で期待しているのが放送の送出系である。性能基準に見合う現実的な伝送容量という制限のあるこうした用途の場合、PCM 方式では弱音部分の 16 ビットデータは大半が無駄になっていることになり、伝送効率面での本質的な弱点となる。DSTL5501 エンコーダ及び 5502 デコーダは AC-2 コーディングを使ってスタジオ・送信機間のデジタルリンクを 950MHz 帯域で実現するシステムで、従来の FM リンクに置き換えて、また混在させての使用も可能である。チャンネルスペーシングわずか 250kHz で、回線 C/N 比の劣化に対してもプログラムの S/N 比が維持されるデジタル伝送の利点が大きい。

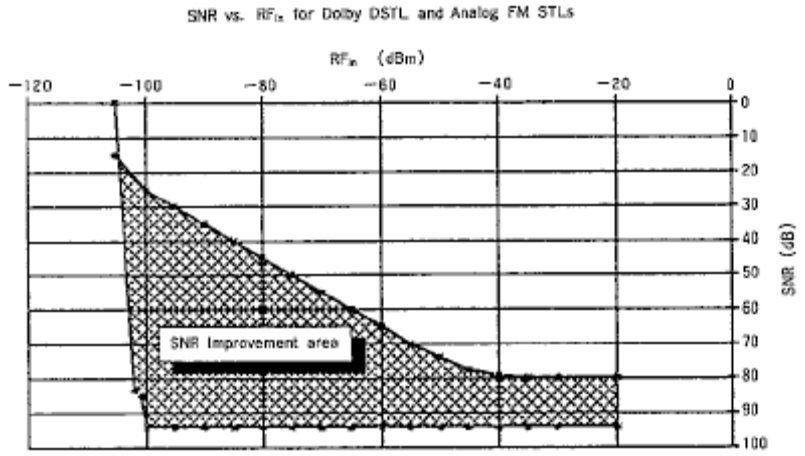

図 4　DSTL の S/N 改善効果

SX-20

この製品は米国 Antex 社がドルビー社より AC-2 技術のライセンス許諾を受けて開発したサウンドボードである。現在 PC の世界でも Windows を軸にマルチメディア化が進められているが、通常搭載される AD/DA ボードはせいぜい 20kHz サンプリングで 8 ビット程度の PCM 処理を行なうものが多い。データレートをこれ以上欲張ることも困難で、性能的に会話用程度のグレードで妥協しているのが実状である。こうした用途に AC-2 を採用することにより、20-20kHz の帯域幅でダイナミックレンジ 90dB 以上という業務用としても通用するだけの性能が初めて登場したことになる。

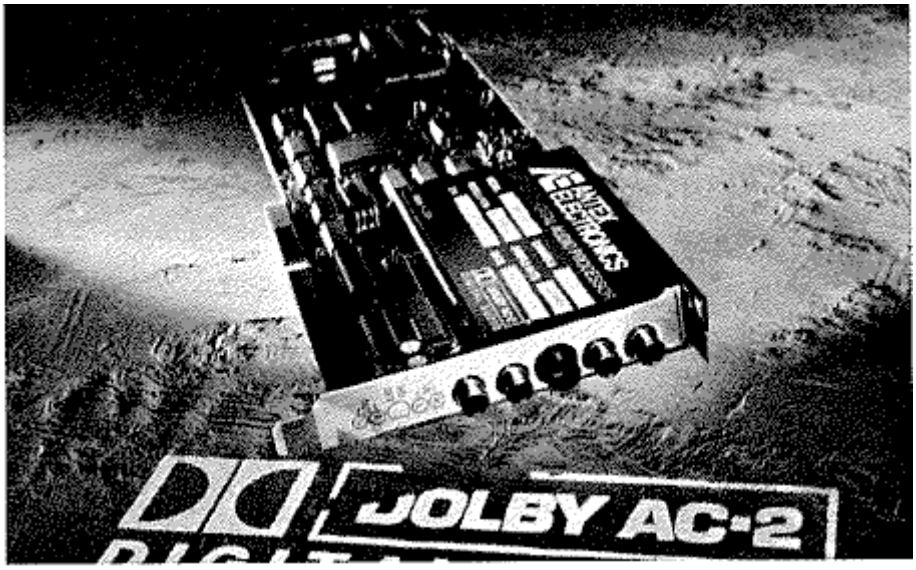

写真2　SX-20サウンドボード

DA-10

SR・D デジタル音声映画システムの劇場及びダビングスタジオ用デコーダである。システムは当然ながら 2 チャンネル AC-2 に対して、5.1 チャンネル AC-3 フォーマットとなっており、今年 7 月米国での『バットマン・リターンズ』封切り以来、作品制作に合わせて、着々と劇場への設備が行なわれている。国内でこうしたシステムを体験できる B も近そうである。

追記

以上、ドルビーAC-2 高性能オーディオ・コーディング方式について概略説明させていただいたが、本稿の中核となっている技術解説の部分はドルビー研究所のスティーブ・フォーシェイによる『Psychoacoustics and Low Bit-Rate Audio Coding』に基づいており、ここではそれを訳出して、アレンジを加えながらまとめてある。低ビットレート・コーディングの基本を理解いただく上で若干でも役に立てばと願う。

Psychoacoustics and Low Bit-Rate Digital Audio Coding

Recent advancements in low bit-rate audio coding technology are based on application of advanced human auditory system models and the underlying perceptual limitations of the ear. The best low bit-rate audio coders deliver sound quality comparable to digital audio compact disc (CD) at a data rate of 128 kbits/second/channel --about one-fifth the data rate of CD. Digital signal processing (DSP) and custom large scale integrated (LSI) circuit devices provide the required computational horsepower and economical means of coder implementation for professional and consumer applications in telecommunications, broadcasting, high-definition television, optical and magnetic recording, and computer multimedia products.

A primary goal in development of low bit-rate audio coders is to provide for cost $ effective transmission and storage of high-quality digital audio. In contrast to speech coding techniques where intelligibility is the key goal and assumptions can made regarding the limited variety of sounds emanating from the vocal tract, low bit-rate audio coders must be designed to work with an unlimited variety of natural and synthesized sounds, with no loss of fidelity. The common element in perception of coded sounds is the human ear, and low bit-rate audio coders are designed to reduce data rate by exploiting perceptual limitations in the human auditory response. Despite the complexities involved and incomplete knowledge as to exactly how the ear responds to complex audio signals, audio coder developers in the US, Europe and Japan have recently made substantial gains in development of mathematical models which characterize auditory limitations and provide the foundation for advanced low bit-rate audio coder design.

The critical-band concept [H. Fletcher, 1940] and the psychoacoustic principles of auditory masking are fundamental to design of effective low bit-rate audio coders.

Auditory masking describes the phenomenon whereby a loud signal tends to “mask” or hide the presence of other quiet signals nearby in frequency. Masking is a consequence of an increase in the ear’s threshold of perception in the frequency range of the loud signal which leaves the ear “deaf” to quieter signals at or near the same frequency.

Results of experiments at low signal levels reveal that a quiet signal which is masked by a louder tone nearby in frequency remains inaudible until the frequency spacing between them exceeds a certain threshold bandwidth, or so called critical-band spacing.

The critical-band model of the ear as a parallel bank of narrow band filters was developed as a means of conceptualizing measured variations in threshold bandwidth or “masking selectivity” as a function of frequency. In the model, notional critical-band filters are approximately 100 Hz in width below 500 Hz, and are of constant fractional bandwidth, i.e. one-fifth of center frequency, above 500 Hz. This model serves as a measure of the

minimum frequency selectivity required to take maximum advantage of the ear's masking characteristics. The important principle to audio coder developers is that the ear functions much like a 25 band real-time spectrum analyzer with bandwidths and sensitivity thresholds that vary somewhat over the 20 Hz to 20 kHz frequency range. Single tone masking experiments indicate that masking effects are minimal within the first 30 dB above the threshold of hearing, i.e. near the quietest sound levels the ear is capable of perceiving. At progressively louder levels, however, masking occurs over a broader frequency range encompassing an increasing number of critical bands, particularly in the frequency range above the masking signal.

Although the body of published data on masking is derived largely from experiments involving sinewaves and narrow-band noise, these data represent applicable upper limits on the thresholds of audibility with more complex audio signals, and are therefore relevant to audio coder design. Single-tone masking curves are well documented in the literature and are beyond the scope of this overview. However, it is useful to identify masking trends for audio signals with predominantly low, middle or high frequency content. Loud, low frequency signals effectively mask the presence of quieter low frequency signals and provide a masking effect which broadens into the mid-frequency range as signal loudness increases. Loud, mid-frequency signals best mask quieter mid and upper-frequency signals; however, this masking effect falls off rapidly just below the frequency range of the masking signal. Loud high-frequency signals effectively mask quieter high frequency signals, but provide very little masking of middle frequencies and no masking at low frequencies. The exact degree of masking is a complex function of the amplitude and distribution of the frequency components of the audio signal, and much remains to be learned in quantifying these complex masking effects.

An additional form of masking that must be considered in optimized low bit-rate audio coder design is that provided by transient signals. Under steady state signal conditions the frequency resolution of the ear is excellent, but it takes the ear a finite time to “tune in” to signal changes, thus implying inherent limitations in time resolution. However, actual measurements in the key time interval just prior to the onset of high level transient test signals confirm time resolution of under 5 milliseconds. Interestingly, masking of quieter signals can occur before, during and after the occurrence of a transient signal. Pre-temporal masking, i.e. that which occurs just prior to the transient, is strongest up to 10 milliseconds before the transient. The masking effect understandably, is strongest during the transient, and falls off over a period of 50-200 milliseconds thereafter.

The fundamental process of low bit-rate coding includes generation of a frequency-domain representation of the audio signal, variable quantization of the signal's

frequency components to a reduced accuracy based on an auditory masking model, allocation of bits to meet the varying demands of the quantizer, and re-synthesis of an approximation of the original time-domain waveform following transmission or storage of the coded data. Generation of the frequency-domain representation of the audio signal is accomplished through use of a multi-frequency band filter bank. Two different frequency division techniques, one based on the discrete Fourier transform [1], and one based on polyphase digital filters [2] have emerged as popular methods. DSP and LSI technology are employed to implement the filter bank, frequency analysis, masking threshold calculation, quantizer and bit-allocation functions.

Audio signals consist of nearly stationary signals and transients changing rapidly with time. Signals which change slowly in time are best coded using a filter bank with a high degree of frequency selectivity such that the spectrum of coding errors may be confined to the spectral region of the signal, and masking may be exploited to best advantage. Transient signals, however, are best coded using a filter bank which has time resolution equal that of the ear, thus avoiding coding errors which spread in time beyond the audibility limits set by the ear's temporal masking characteristics. As excellent frequency selectivity and short time resolution are mutually exclusive requirements, filter bank design can involve a) a compromise between time and frequency resolution with sufficient bit-rate allocated to meet temporal and spectral masking constraints, or b) a filter bank with time-varying optimization for either time or frequency resolution depending on the characteristics of the signal to be coded. Both techniques are employed in low bit-rate audio coders.

A conceptual block diagram of a multi-frequency band low data-rate audio coder is illustrated in Figure 1. A sampled and quantized time domain input signal consisting of low, middle and high frequency components A, B and C is converted to a frequency-domain representation using an appropriate filter bank technique with critical-band frequency resolution. The frequency-domain representation of the signal is shown in Figure 1a. Once the frequency components of the audio signal are identified, an estimate of the masking thresholds is made on a band-by-band basis by direct calculation, or by comparison with a pre-programmed model of the ear. Having established the masking thresholds based on the loudest signal components present in each of the frequency bands, other nearby signal components are analyzed to determine whether the loudest signals provide sufficient masking to render quieter signals within that same critical band inaudible. Once this determination is made, the bit-rate reduction portion of the process can be completed. This involves quantizing the frequency components in each of the individual filter bands with sufficient accuracy to keep the quantization noise just below the calculated in-band masking thresholds, and amplitude scaling of the signals to

normalize their peak levels to make optimum use of digital signal processor dynamic range. Note that although a low bit-rate representation of the signal has been created, a wideband noise component is introduced as a result of quantizing the frequency-domain signals. The crosshatched areas in Figure 1b represent the quantization noise added to the signal components, just below the in-band masking thresholds. Based on masking criteria discussed earlier, this wideband noise component would not be effectively masked by the signal components if no further action was taken, and the desired high fidelity would not be achieved. However, an additional process takes place in the decoder after transmission or storage of the coded representation of the signal, where an identical filter bank to that used in the encoder is employed. Quantized frequency-domain data is received by the decoder and passed through the filter bank. This re-filtering process leaves the frequency components of the signal intact, while tightly constraining the unwanted quantizer-introduced noise to a narrow frequency range below the masking threshold as shown in Figure 1c. As long as filter bank frequency selectivity and out-of-band signal rejection are sufficient, masking thresholds may be conservatively applied with sufficient bit rate allocated to keep quantizer noise below audible limits, and the reconstructed time waveform at the decoder output will sound subjectively equivalent to that of the input signal.

The design of high-quality low bit-rate audio coders involves a trade-off between the degree of bit rate reduction and subjective audio quality. Systems currently available achieve near-perceptual transparency for 20 Hz to 20 kHz bandwidth audio signals at a data rate of 128 kbits/second/channel. Coder development work continues towards total transparency at current data rates, and equivalent sound quality at lower data rates.

Steven E. Forshay
11-Aug-92 6:01 PM

Bibliography

1. Marina Bosi and Grant Davidson, “High Quality, Low-Rate Audio Transform Coding for Transmission and Multimedia Applications,” 93rd Convention of the Audio Engineering Society, San Francisco, October 1-4, 1992.

2. G. Stoll and Y. F. Dehery, “High Quality Audio Bit-Rate Reduction System Family for Different Applications,” Proceedings of the IEEE International Conference on Communications, Atlanta, April 1990.