TEMPORAL NOISE SHAPING, QUANTIZATION AND CODING METHODS IN PERCEPTUAL AUDIO CODING: A TUTORIAL INTRODUCTION

JÜRGEN HERRE[1]

[1]*Fraunhofer Institute for Integrated Circuits FhG-IIS A, Erlangen, Germany*
hrr@iis.fhg.de

Perceptual audio coding has become an important key technology for many types of multimedia services these days. This paper provides a brief tutorial introduction into a number of issues as they arise in today's low bitrate audio coders. After discussing the Temporal Noise Shaping technology in the first part of this paper, the second part will focus on the large number of possible choices for the quantization and coding methods for perceptual audio coding along with examples of real-world systems using these approaches.

INTRODUCTION

Perceptual audio coding has become an important key technology for many types of multimedia services including audio playback and storage.

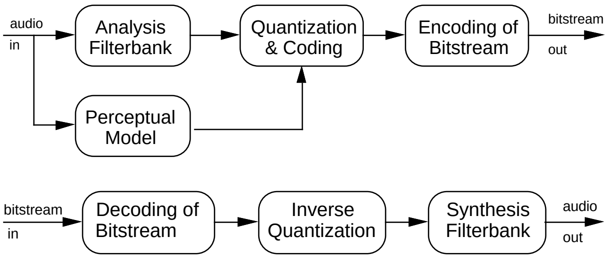

Figure 1: Generic structure of a perceptual audio encoder (above) and decoder (below).

Generally, the well-known generic structure of a perceptual audio coder for monophonic audio signals can be described in the following way (see Figure 1):

- The input samples are mapped into a subsampled spectral representation using an analysis filterbank.
- Using a perceptual model the signal's frequency and time dependent masking threshold is estimated. This gives the maximum coding error that can be introduced into the audio signal while still maintaining perceptually unimpaired signal quality.
- The spectral values are then quantized and coded according to requirements derived from the masking threshold estimate. In this way, the quantization noise is hidden ("masked") by the respective transmitted signal as far as possible and perceptibility of the coding error is minimized.
- Finally, all relevant information (i.e. coded spectral values and additional side information) is packed into a bitstream and transmitted to the decoder.

Accordingly, the order of processing appears reversed in the corresponding decoder:

- The bitstream is decoded and parsed into coded spectral data and side information.
- The inverse quantization of the quantized spectral values is carried out.
- The spectral values are mapped back into a time domain representation using a synthesis filterbank.

This paper provides a tutorial overview over certain aspects of perceptual audio coding and is structured as follows: The first part of the paper will address the general issue of the temporal masking problem in perceptual audio coding and describe the so-called Temporal Noise Shaping (TNS) technology. In particular, the principle and theoretical background of the TNS approach is discussed along with its integration into a perceptual audio codec.

In the second part of the paper, the fundamental issues of quantization and coding in a perceptual audio coder are considered and the most important technical approaches are contrasted. Examples will be given for the use of the discussed techniques in standardized or industry standard codecs. Finally, a discussion of encoding strategies will conclude the paper.

1. TEMPORAL NOISE SHAPING (TNS)

The *Temporal Noise Shaping* (TNS) technique represents a rather novel concept in perceptual audio coding and was first introduced in 1996 [11]. As outlined below, it can be considered an extension of the basic scheme of a perceptual coder (Figure 1), inserting a new, optional processing step between the filterbank and the quantization/coding stage, thus further optimizing the coder's performance.

1.1 The Temporal Masking Problem

The approach is motivated by the fact that, despite the advanced state of today's perceptual audio coders, the handling of transient and pitched input signals still presents a major challenge. This is mainly due to the temporal aspect of masking: In order to achieve perceptually transparent coding quality the quantization noise must not exceed the time-dependent masking threshold.

In practice, this requirement is not easy to meet for perceptual coders because using a spectral signal decomposition for quantization and coding implies that a quantization error introduced in this domain will be spread out in time after reconstruction by the synthesis filterbank (time/frequency uncertainty principle). For commonly used filterbank designs (e.g. a 1024 lines Modified Discrete Cosine Transform, MDCT [22]) this means that the quantization noise may be spread over a period of more than 40 milliseconds (assuming a sampling rate of 48 kHz). This will lead to problems when the signal to be coded contains strong signal components only in parts of the analysis filterbank window, i. e. for transient signals. In particular, quantization noise is spread *before* the onsets of the signal and, in extreme cases, may even exceed the original signal components in level during certain time intervals. Such a constellation is traditionally known as a "pre-echo phenomenon" [10].

Due to the properties of the human auditory system, such "pre-echoes" are masked only if no significant amount of the coding noise is present longer than ca. 2 ms before the onset of the signal. Otherwise the coding noise will be perceived as a pre-echo artifact, i.e. a short noise-like event preceding the signal onset. In order to avoid such artifacts care has to be taken to maintain appropriate temporal properties of the quantization noise such that it will still satisfy the conditions for temporal masking. This *temporal noise shaping problem* has traditionally made it difficult to achieve a good perceptual signal quality at low bitrates for transient signals, such as castanets, glockenspiel, triangle or certain types of speech signals.

The Temporal Noise Shaping technique permits the coder to exercise some control over the temporal fine structure of the quantization noise even *within* each filterbank window and in this way addresses the above problems. The following section is dedicated to the discussion of the principles underlying this technique.

1.2 Principle of TNS

Basically, the Temporal Noise Shaping (TNS) approach is based on two main considerations, namely

1 . Consideration of the time/frequency duality between spectral envelope and (squared) Hilbert envelope and
2. Shaping of quantization noise spectra by means of open-loop predictive coding

1.2.1 Time / Frequency Duality Considerations

The concept of TNS is based upon the dual of the standard LPC analysis paradigm. It is well-known that signals with an "un-flat" spectrum can be coded efficiently either by directly coding spectral values ("transform coding") or by applying predictive coding methods to the time signal [14]. Consequently, the corresponding dual statement relates to the coding of signals with an "un-flat" time structure, i.e. transient signals. Efficient coding of transient signals can thus be achieved by either directly coding time domain values or by employing *predictive coding methods to the spectral data* by carrying out a prediction across frequency. In fact, it can be shown that, due to the duality between time and frequency, the amount of "prediction gain" (i.e. reduction of residual energy) reached is determined by the "unflatness" of the signal's temporal envelope in the same fashion that the Spectral Flatness Measure (SFM) is a measure of the reduction of residual energy available by LPC prediction [14].

A more rigorous derivation of these properties was published in [11] showing that the squared Hilbert envelope of a signal and the power spectral density constitute dual aspects in time and frequency domain.

1.2.2 Noise Shaping by Predictive Coding

Figure 2 illustrates the structure of a traditional open-loop predictive coding system. The correlation between subsequent input samples is exploited by quantizing / coding the prediction error based on the unquantized input samples. This scheme is also known as "forward prediction" (D*PCM) as opposed to the more widely used "backward prediction" approach which comprises a prediction based on previously quantized values [14].

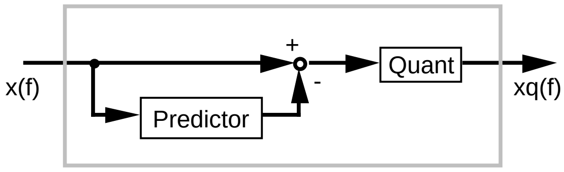

Figure 2: Open-loop predictive encoding (D*PCM).

If D*PCM is applied to *a time signal*, the quantization error in the final decoded signal is known to be *adapted in its Power Spectral Density* (PSD) to the PSD of the input signal [14]. Combining this observation with the time/frequency duality noted above, the following dual statement can be derived: If such predictive coding is *applied to spectral data over frequency*, the *temporal shape of the quantization error* signal will appear adapted to the temporal shape of the input signal at the output of the decoder.

As a consequence, this effect can be used to put the quantization noise of a filterbank-based coder under the actual signal and in this way avoids problems of temporal (un)masking, both in transient or pitched signals. This type of predictive coding of spectral data is referred to as the “Temporal Noise Shaping” (TNS) approach.

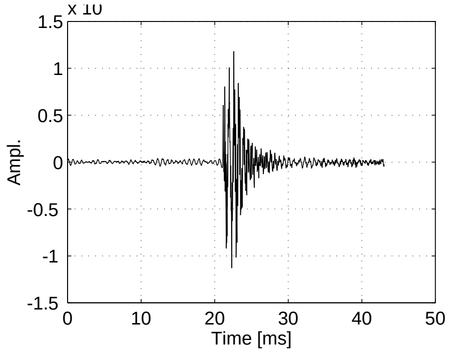

Figure 3: Transient signal (castanets, uncoded).

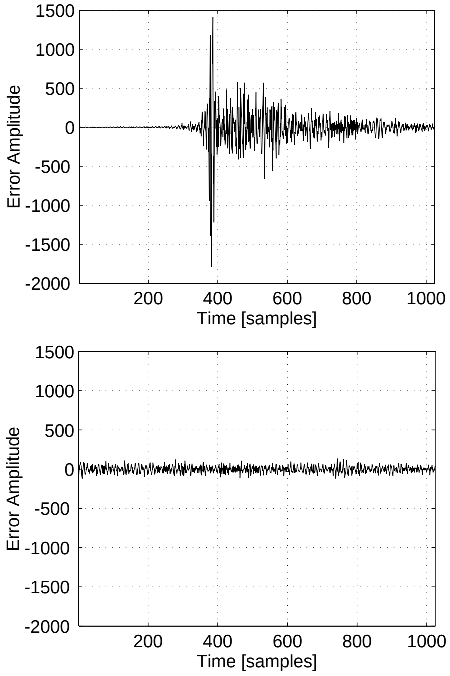

Figure 4: Coding noise in decoded castanets signal with (above) and without (below) TNS.

Figures 3 and 4 illustrate the noise shaping effect for a transient signal (castanets). As expected from above considerations, the temporal shape of the coding noise is adapted to the envelope of the input signal by TNS whereas in the case of the standard processing (no TNS) the quantization noise is distributed almost uniformly over time.

A more extensive review of applying these dualities for the purpose of noise shaping, joint stereo coding and other purposes can be found in [12].

1.3 TNS – A Predictive Coding Method

The TNS predictive encoding / decoding process over frequency can easily be realized by adding one block to the standard structure of a generic perceptual encoder and decoder. This is shown in Figure 5. An additional block, “TNS Filtering”, is inserted after the analysis filterbank performing an in-place filtering operation on the spectral values, i.e. replacing the target spectral coefficients (set of spectral coefficients to which TNS should be applied) with the prediction residual. This is symbolized by a rotating switch circuitry in the figure. Both sliding in the order of increasing and decreasing frequency is possible.

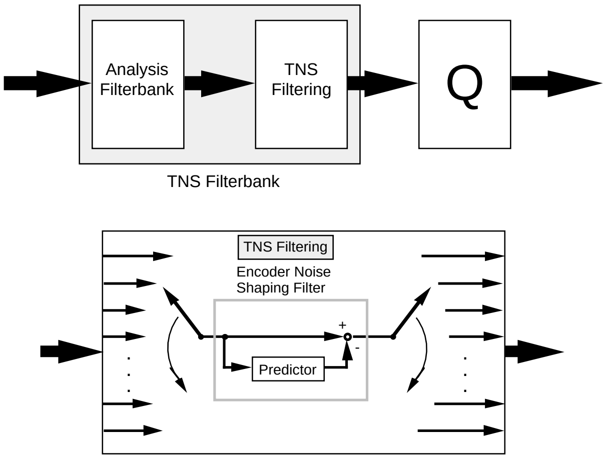

Figure 5: TNS Processing in the encoder.

Similarly, the TNS decoding process is done by inserting an additional block, "Inverse TNS Filtering", immediately before the synthesis filterbank (see Figure 6). An inverse in-place filtering operation is performed on the residual spectral values such that the target spectral coefficients are replaced with the decoded spectral coefficients by means of the inverse prediction (all-pole) filter.

The TNS operation is signaled to the decoder via a dedicated part of the side information that includes a TNS on/off flag and the prediction filter data.

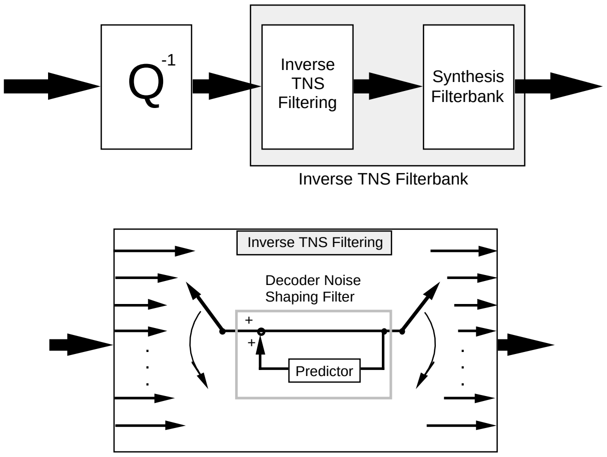

Figure 6: TNS Processing in the decoder.

While the interpretation of the TNS approach was based on considerations of predictive coding methods, it is most instructive to look upon the resulting coding scheme from a filterbank point of view.

1.4 TNS – A Continuously Signal Adaptive Filterbank

In fact, it is possible to interpret the combination of filterbank and prediction filter as a composite filterbank (the "Temporal Noise Shaping Filterbank", see Figures 5 and 6) with a number of interesting properties [13].

In contrast to the well-known block switching approach [26], the Temporal Noise Shaping filterbank allows a continuous adaptation to the properties of the input signal in the following way:

- For signals with a considerable correlation between adjacent spectral coefficients (i.e. for signals with a very "unflat" temporal envelope) the prediction filter will combine (convolve) these coefficients to calculate the prediction residual.
- In this way, frequency resolution will decrease and is traded adaptively in favor of temporal resolution. Note that the increased temporal resolution of the filterbank is not represented by a number of timely subsequent spectral coefficients but by a multitude of coefficients of the same time instant corresponding to largely overlapping (widened) frequency bins.

In this way, the frequency (and time) resolution is adjusted adaptively to the input signal. This enables the interpretation of the combination of filterbank and adaptive prediction filter as a *continuously adaptive filterbank* as opposed to the classic "switched filterbank" paradigm. In fact, this type of adaptive filterbank dynamically provides a continuum in its behavior between a high-resolution filterbank (for stationary signals) and a low-resolution filterbank (for transient signals) and therefore approaches the requirements for the optimum filterbank for a given input signal.

1.5 TNS and Time Domain Aliasing

So far the discussion of Temporal Noise Shaping was based on the notion of Fourier transform (and Discrete Fourier Transform, DFT, in the case of discrete spectral coefficients). In practice, the Modified Discrete Cosine Transform (MDCT) is preferred over the DFT or Discrete Cosine Transform (DCT) in a modern filterbank-based coder for the reasons that it is both critically sampled and delivers excellent coding efficiency.

Extending the considerations beyond DFT, it can be shown that the TNS filterbank provides a straightforward temporal noise shaping effect also for the known classic orthogonal block transforms, like DCT or Discrete Sine Transform (DST). Furthermore, if the perceptual coder uses a critically subsampled filterbank with overlapping windows (e.g. an MDCT or any other filterbank based on Time Domain Aliasing Cancellation TDAC [22]), the resulting temporal noise shaping is also subject to the time domain aliasing effects inherent in this filterbank. For example, in the case of a MDCT one mirroring (aliasing) operation per window half takes place and the quantization noise appears mirrored (aliased) within the left and the right half of the window after decoding, respectively. Since the final filterbank output is obtained by applying a synthesis window to the output of each inverse transform and performing an overlap-add of these data segments, the undesired aliased components are attenuated depending on the shape of the analysis-synthesis window pair. Thus it is advantageous to choose a filterbank window that exhibits only a small overlap between subsequent blocks such that the temporal aliasing effect is minimized.

This approach was chosen for the MPEG-4 Version 2 AAC-derived Low Delay coder [5] [23]. Interestingly, this coder– unlike most of today's codecs - does not employ block switching methods but predominantly relies on TNS to address the temporal masking problem thus avoiding the structural complication associated with block switching. A second filterbank window with low overlap is available to minimize temporal aliasing and in this way optimize TNS performance for transient signal types. Figure 7 demonstrates this effect for a castanets signal.

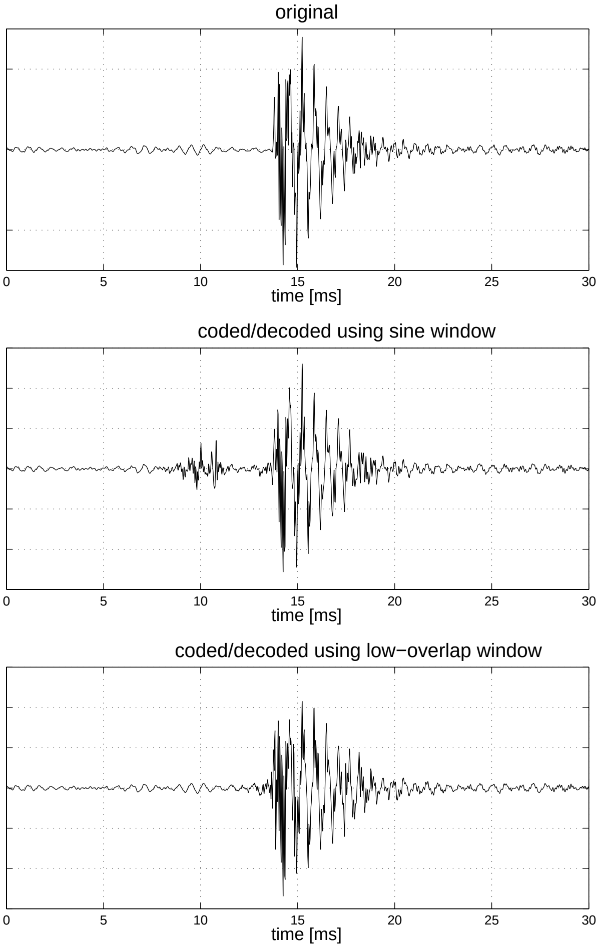

Figure 7: Reduction of TNS temporal aliasing effect by using a low overlap window
(Low Delay AAC Coder: 64 kbps; castanets signal)

1.6 Using TNS in a Perceptual Audio Coder

In general, TNS offers the following benefits for a perceptual audio coder:

- It permits for a better encoding of ”pitch-based” signals, such as speech which consist of a pseudo-stationary series of impulse-like signals without penalty in coding efficiency. Figure 8 illustrates this application. Here the original waveform (top plot) is shown together with the introduced coding noise with (middle plot) and without (bottom plot) TNS processing, respectively. The pitched structure of the speech signal is clearly visible as well as the concentration of the coding noise around each glottal pulse in the case of TNS. In contrast, the standard processing (lower plot) shows only a much smaller degree of noise shaping so that unmasking is more likely to occur.
- The method reduces the peak bit demand of the coder for transient signal segments by reducing the required pre-echo protections for such signals. As a side effect, the coder can stay longer in the preferred ”long block” mode so that use of the less efficient ”short block” mode can be reduced.

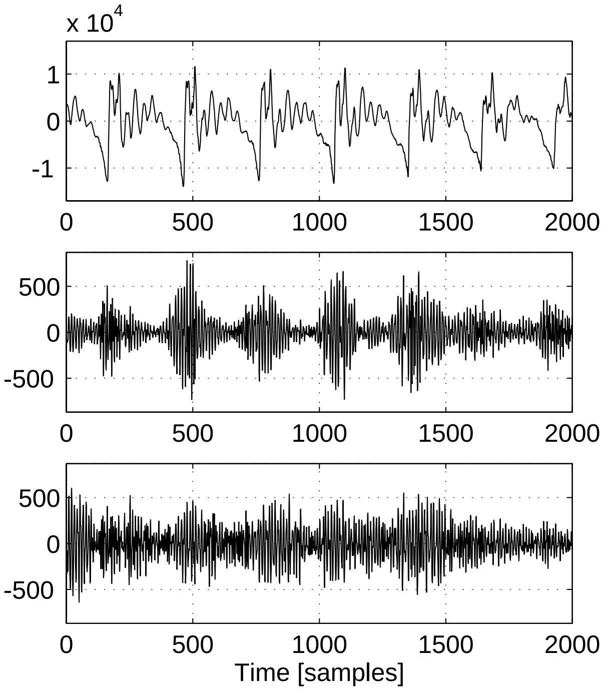

Figure 8: Original signal and coding noise with and without TNS (from top to bottom; speech excerpt)

- The technique can be combined with other methods for addressing the temporal noise shaping problem, such as block switching. Using temporal noise shaping it may, however, be possible to omit the need for a second coder mode (short block mode) leading to a simplified encoder / decoder structure.
- Since TNS processing can be applied either for the entire spectrum or for only part of the spectrum, the time-domain noise control can be applied in any necessary frequency-dependent fashion. In particular, it is possible to use several filters operating on distinct frequency (coefficient) regions, or to provide no TNS processing at some frequencies.

The first widely used audio codec exploiting TNS was the MPEG-2 AAC coding system [3] [7]. Subsequently, the MPEG-4 General Audio coding [4] as well as the MPEG-4 version 2 Low Delay AAC coder [5] [23] employed this technique, the latter even without additional filterbank block switching.

2. QUANTIZATION AND CODING

While Temporal Noise Shaping addresses very specific aspects of the coder's performance, quantization and coding are basic building blocks which are essential for the function of all coders. In general, the goal of these two steps is to achieve a representation of the spectral data which is both as compact as possible (low bitrate) and at the same time introduces as little perceptible distortion into the signal as possible (high perceptual quality). The performance of the coder's quantization / coding kernel is characterized by its ability to reconcile these conflicting requirements.

Although this is not required from a principal point of view, the quantization / coding process is carried out in two separate steps by the vast majority of common audio codecs. Firstly, a quantization step maps the spectral coefficients to quantized values and indices. Subsequently, these index values are coded compactly by a so-called noiseless coding kernel for subsequent transmission. Accordingly, both aspects will be addressed separately in the next subsections, followed by a short discussion of joint quantization / coding techniques. The use of the discussed techniques will be illustrated by examining a number of standard or industry standard coding schemes, including the MPEG-1/2 Layer I, II and III coder [1] [2] [6], MPEG-2 Advanced Audio Coding (AAC) [3] [7], MPEG-4 audio coding [4], Sony's ATRAC coder [8] and Dolby's AC-3 system [9].

2.1 Quantization

In the quantization stage, a reduction in the precision of the representation of the input signal is carried out by mapping the set of spectral coefficients originating from the analysis filterbank into a finite set of quantized values and corresponding index values. This enables a reduction in transmission rate at the expense of introducing quantization distortion and is the general mechanism for exploiting irrelevancy in a perceptual coder.

As a fundamental difference compared to other coder types, the precision of quantization (quantization step size) is controlled carefully in a perceptual coder according to the time and frequency dependent masking threshold which is estimated by the coder's perceptual model. Consequently, one of the requirements for the quantizer stage is a large amount of flexibility regarding the quantization precision it can deliver. Frequently, transparent coding of signals may require local Signal-to-Noise-Ratio values ranging from 0 dB (transmission of spectral coefficients not necessary) to 30 dB or even higher (e.g. in critical tonal or pre-echo conditions). Since this control is done in a frequency selective way, the resulting quantization distortion after decoding of the signal appears shaped according to the perceptual requirements (*noise shaping*).

When trying to achieve a precise control of the resulting coding distortion in the final reconstructed signal it is important to consider the spread of distortion in both time and frequency due to the used synthesis system (i.e. synthesis filterbank). Thus, directly applying the raw thresholds calculated by the psychoacoustic model as thresholds for quantization of the spectral coefficients is not always sufficient. Instead, these raw values need to be corrected for effects of the synthesis system such that, after decoding, the desired distortion criteria are met.

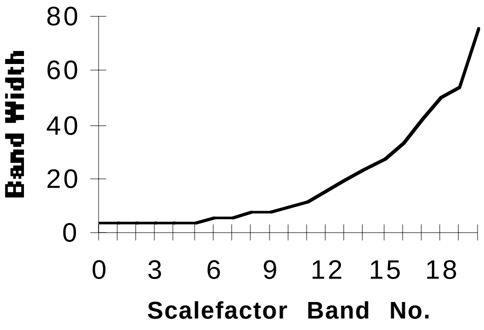

Figure 9: MPEG-1 Layer 3 scalefactor band size (for fs=44.1 kHz / long blocks).

Usually, the quantization step size applied in the encoder needs to be communicated to the decoder by including some amount of side information into the bitstream. To keep the amount of side information within reasonable bounds, the same quantizer resolution is employed for groups of spectral coefficients which share this part of the side information (sometimes called *scalefactor bands*). These groups of coefficients may consist of spectral values that are subsequent in time or adjacent in frequency depending on the time / frequency resolution of the filterbank. For example, while 12 subsequent output values of the polyphase filterbank are grouped in MPEG-1/2 Layers I and II, MPEG-1/2 Layer III groups adjacent spectral "lines". This grouping scheme can be used to adjust the frequency resolution of the coder noise shaping to the resolution of the human auditory system (e.g. related to BARK [20] or ERB [21] scales). As an example, Figure 9 shows the size of the scalefactor band groups of MPEG-1 Layer III, expressed in number of spectral lines.

2.1.1 Scalar Quantization

Most of today's perceptual audio coders use scalar quantization of the spectral coefficients, i.e. each spectral component is quantized separately. The advantage of this simple approach can be seen both in its low computational complexity and the fact that scalar quantizers can be easily scaled in resolution over a wide range.

Most simple in its implementation, uniform scalar quantizers have gained wide use in coding system, such as MPEG-1/2 Layers I and II, ATRAC and AC-3. More sophisticated in structure, non-uniform quantization is used e.g. in MPEG-1/2 Layer III and MPEG-2/4 AAC coders by employing a power law (x^0.75) quantizer. While uniform quantization will on average produce the same amount of quantization noise for each coefficient of the scalefactor band (Q^2/12, denoting the quantizer step size by Q), a power-law quantizer tends to distribute the overall distortion towards spectral coefficients with large amplitudes where the probability of perceptual masking is higher. In this way, noise shaping is carried out implicitly even within each scalefactor band. Another benefit of the compressive power-law quantizer is that, by reducing the dynamics of the spectral coefficients, it effectively helps to reduce the amount of quantizer step size information and the size of Huffman code tables (if Huffman coding is used subsequently).

Typically, two different approaches are used as scalar quantization schemes: Firstly, the group of spectral coefficients is normalized by a common multiplier ("scalefactor") to match the operation range of the quantizer and is then quantized using the same quantizer resolution (number of bits per sample). Both the multiplier and the quantizer resolution are transmitted to the decoder as side information. This scheme is traditionally known as "block companding" or "block floating point" [14] and is used widely in low complexity coders, like MPEG-1/2 Layers I and II, ATRAC and AC-3. Accordingly, different terminology is used for the groups of spectral coefficients ("scalefactor bands", "block floating units"), the common multiplier ("scalefactor", "exponent"), the quantized values ("mantissa") and the used quantizer resolution ("bit allocation information").

The second widely adopted quantization scheme is employed in more sophisticated coders using Huffman coding of quantized coefficients (e.g. MPEG-1/2 Layer III, MPEG-2 AAC, see below). Depending on a "scalefactor", the spectral values are scaled and subsequently quantized by a fixed quantizer. In this way, the effective quantizer resolution is controlled by the scalefactor: Large scaling values will result in a fine quantization (and big quantized values to be coded subsequently), and small scaling values will produce the converse result. This scheme requires only one information (i.e. the scalefactor) to be transmitted to the decoder.

It is important to note that the same terms (e.g. "scalefactors") are often used with different meanings in the context of both quantization scenarios: While Layer II scalefactors give a good indication about the signal amplitude in a particular band, Layer III scalefactors do not allow such conclusions since the actual magnitude also depends on the magnitude of the quantized coefficients.

2.1.2 Vector Quantization

In contrast to the scalar quantization, vector quantization (VQ) performs a joint quantization and coding of vectors of input values. Since VQ combines the steps of quantization and coding, VQ-based coding will be discussed in section 2.3 (joint quantization / coding techniques).

2.2 Noiseless Coding

The objective of the noiseless coding stage in a perceptual audio coder is to achieve a further gain in required data rate by reduction of redundancy in the representation of the transmitted data. This is done by a lossless packing of quantized spectral data exploiting statistical dependencies and other properties.

Again, there are a number of possible and common approaches: In the case of the straight-forward block companding scenario, no additional noiseless coding is applied usually, i.e. the quantized data are transmitted as simple PCM codes.

In some schemes, a simple packing mechanism helps to avoid excessive overhead caused by inefficient usage of the PCM code range. For example, transmission of 3 values originating from a 5-level quantizer as separate code words would require 3x3=9 bits whereas combination of this data into a single codeword will only require 7 bits. In MPEG-1/2 Layer II this mechanism is known as "grouping" and requires just very moderate additional computational effort.

Higher coding gain can be achieved by using entropy coding techniques for noiseless coding. In particular, Huffman coding is a popular technique for exploiting redundancy in the quantized spectral data. Additional coding gain can be achieved by using multi-dimensional Huffman coding, i.e. combining several spectral coefficients into a vector which is coded with one Huffman code word, thus exploiting the joint statistics of the vector components. Optionally, the sign information can be coded as part of the Huffman code or separately. To illustrate the structure of common noiseless coding kernels, we will briefly look into the respective parts of the MPEG-1/2 Layer III and MPEG-2 AAC coders.

2.2.1 MPEG-1/2 Layer III

The Layer III noiseless coding strategy is best explained by proceeding from high frequency to low frequency spectral components. Since there is typically a large number of "zero" coefficients at the upper end, these coefficients are transmitted by run-length coding, i.e. indicating the number of zero spectral coefficients. Next, a region of spectral values is coded with absolute values not exceeding one. This is done using one of two alternate 4-dimensional Huffman code books. The remaining section of coefficients contains the most significant part of the spectral information and can be partitioned into up to three *regions*. A different Huffman code book can be chosen for each region from a number of 2-dimensional Huffman code books with different sizes and maximum absolute values. Spectral coefficients with an absolute value exceeding 15 are coded by using *escape code books*.

2.2.1 MPEG-2 AAC

A further increase in coding efficiency and versatility is achieved by using the AAC noiseless coding kernel [15]. Compared to the Layer III kernel, AAC provides more 4-dimensional tables and an extremely flexible mechanism called "sectioning". This scheme allows to partition the set of quantized spectral values into arbitrary groups of scalefactor bands ("sections") each of which is assigned a specific Huffman code table (Figure 10). Accordingly, for each section the number of included scalefactor bands and the assigned code book number is transmitted to the decoder. As a result, efficient coding can be achieved even in cases when the local statistics of the spectral coefficients are changing frequently within the coded set of coefficients (e.g. due to the frequency-selective use of joint stereo coding techniques).

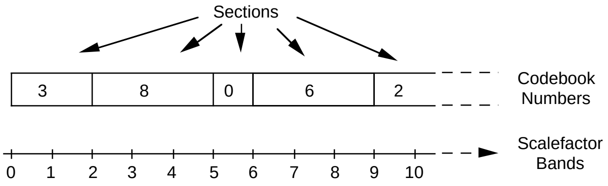

Figure 10: MPEG-2 AAC sectioning.

Beyond Huffman coding, also other entropy coding techniques, like arithmetic coding, have been used for efficient coding of the quantized spectral data. As an example, Bit-Sliced Arithmetic Coding (BSAC) has been proposed to provide fine-grain bitrate scalability of compressed audio [16]. This is achieved by reordering the binary values of the quantized spectral coefficients into "slices" of bits, assembling vectors of MSBs, LSBs and intermediate bits which are subsequently coded by an arithmetic coder. BSAC is currently part of the MPEG-4 version 2 committee draft [5].

2.3 Joint Quantization / Coding Techniques

Joint quantization / coding techniques combine both the reduction of precision by means of quantization and utilization of redundancy in one single step. In particular, the vector quantization (VQ) technique is a multi-dimensional generalization of the well-known Lloyd-Max quantizer [14] and is able to exploit both statistical redundancy of a single quantized component as well as between each of the vector components.

As is well-known from theory, vector quantization offers excellent coding efficiency even at very low data rates (far below 1 bit/sample) when scalar quantization performance degrades rapidly. On the other hand, the high variability of required coding precision, as dictated by perceptual considerations, is hard to satisfy using a table-based VQ scheme. In particular, transparent or near-transparent coding of e.g. tonal signal components will require a very high coding precision (e.g. 30 dB local SNR) which would demand for extremely large VQ tables and a computationally complex search process. As a consequence, the primary area of application for current VQ-based audio coding schemes can be seen in low/intermediate quality coding at very low data rates. A recent example for such a scheme is the Transform-domain Weighted Interleave Vector Quantization (TwinVQ) coder [17] which has been adopted by MPEG-4 as a very low bitrate kernel for scalable audio coding down to 6 kbit/s [18].

The TwinVQ coder performs a quantization of the spectral coefficients in two steps: In a first step the spectral coefficients are normalized to a specified target range and are then quantized by means of a weighted vector quantization (VQ) process. The spectral normalization process includes a linear predictive coding (LPC) spectral envelope estimation scheme, a periodic component extraction scheme, a Bark-scale spectral estimation scheme, and a power estimation scheme which are carried out sequentially. As a result, the spectral coefficients are "flattened" and normalized across the frequency axis. The parameters associated to the spectral normalization process are quantized and transmitted as side information.

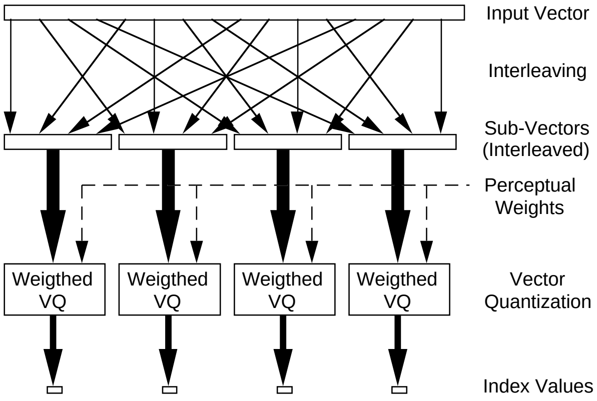

Figure 11: TwinVQ Interleaving and Quantization.

In the subsequent weighted vector quantization process (see Figure 11), the flattened spectral coefficients are interleaved and divided into sub-vectors for vector quantization. For each sub-vector, a weighted distortion measure is applied to the conjugate structure VQ which uses a pair of code books. In this way, perceptual control of the quantization distortion is achieved. The main part of the transmitted information consists of the selected code book indices. Due to the nature of the interleaved vector quantization scheme, no adaptive bit allocation is carried out for individual quantization indices (i.e. an equal amount of bits is spent for each of the quantization indices).

2.4 Overview Over Common Codecs

The following table provides an overview about the basic types of quantization / coding used in standard or industry standard coding schemes.

	Filterbank channels	Quantization	Coding
MPEG-1/2 Layer I / II	32	uniform	block companding
MPEG-1/2 Layer III	576 / 192	non-uniform	Huffman coding
MPEG-2 AAC	1024 / 128	non-uniform	Huffman coding
MPEG-4 TwinVQ	1024 / 128 (960/120)	VQ	VQ
AC-3	256 / 128	uniform	block companding
ATRAC	96 .. 512	uniform	block companding

Note: In the case of ATRAC, the hybrid filterbank (based a two-stage QMF filter and subsequent MDCT processing) allows multiple filterbank configurations with up to 512 filterbank channels with unequal frequency resolution.

3. ENCODING STRATEGIES

Typically, today's perceptual audio coders provide a large amount of flexibility regarding how to encode a particular input signal. Some examples of this flexibility include the choice of the

- quantization noise profile over frequency
- trade-off audio bandwith versus overall distortion
- bitrate and coding mode (independent or joint stereo)
- usage of additional / optional coding tools, such as joint stereo coding, TNS, prediction

In this sense, the encoding strategy is the actual "intelligent" part of the encoding process and determines the resulting audio quality to a significant degree. Consequently, this is the arena for specific implementation know-how ("secrets of audio coding") which can result in major differences between different encoder implementations.

This chapter will discuss some of the possibilities and common strategies regarding the encoding process. In particular, the different approaches for quantization of spectral components under bitrate constraints are outlined and contrasted.

Figure 12 introduces a number of important terms in the context of perceptual coding which are typically evaluated within groups of spectral coefficients ("coder bands") in a coder.

- The Signal-to-Noise-Ratio (SNR) denotes the ratio between signal and quantization noise energy and is a commonly used distortion measure based on a quadratic distance metric. Please note that this measure in itself *does not* allow predictions of the subjective audio quality of the decoded signal. Clearly, reaching a higher local SNR in the encoding process will require a higher number of bits.
- The Noise-to-Mask-Ratio (NMR) is defined as the ratio of the coding distortion energy with respect to the masking threshold and gives an indication for the perceptual audio quality achieved by the coding process. While it is the goal of a perceptual audio coder to achieve values below 0 dB ("transparent coding"), coding of "difficult" input signals at very low bitrates is likely to produce NMR values in excess of this threshold, i.e. a perceptible quality degradation will result from the coding/decoding process.
- The Signal-to-Mask-Ratio (SMR) describes the relation between signal energy and masking threshold in a particular coder band. This parameter significantly determines the number of bits that have to be spent for transparent coding of the input signal. As one extreme case, if the signal energy is below the masking threshold, no spectral data needs to be transmitted at all and there is zero bit demand. A generalization of this concept is called the *Perceptual Entropy* (PE) [24] which provides a theoretical minimum bit demand (entropy) for coding a signal based on a set of perceptual thresholds.

Furthermore, it is important to distinguish between two different coding approaches, namely the *constant quality coding* and the *constant rate coding* scenarios.

3.1 Constant Quality Coding

In order to code an input signal with a pre-selected constant quality, the following steps have to be carried out:

- Determine the time and frequency dependent masking threshold for the signal segment to be coded

- Adjust the quantizer step size in each coder band such that the target distortion (e.g. the masking threshold) is met

Based on this set of quantizer settings, the amount of bits is determined to represent the input signal at the desired degree of precision. Consequently, coding at constant perceptual quality will generally result in a time-varying bitrate according to the spectral and perceptual characteristics of the input signal. An important special case is *coding at threshold* which targets at perceptually transparent coding (without coding headroom) at the lowest possible bitrate (ideally at the Perceptual Entropy). Such a coding system sometimes is referred to as a *threshold simulator* [25].

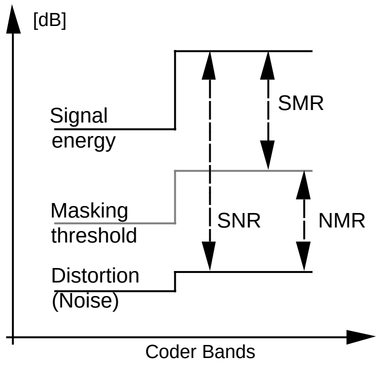

Figure 12: SNR, SMR and NMR.

While constant quality coding is an attractive concept in a perceptual sense, the associated variable data rate restricts the use of this approach to application scenarios which support a variable transmission rate. Some examples for applications of variable rate coding include digital storage applications like compression of multi-channel sound on Digital Versatile Disc (DVD) as well as audio on the Internet.

3.2 Constant Rate Coding

In contrast to *constant quality* coding, the *constant rate* coding approach aims at producing a coded output stream with a fixed bitrate, consequently accepting a time-varying coding quality depending on how demanding an input signal is for the particular coding system.

In general, the constant rate coder can be regarded as today's "default" coder for commercial applications and is a mandatory choice for applications with a fixed rate transmission channel, such as Digital Audio Broadcasting (DAB) or transmission over ISDN channels.

While a quantization strategy for constant quality coding is entirely based on perceptual criteria, constant bitrate coding requires both the consideration of perceptual and rate requirements. More specifically, the goal is the optimization of perceptual quality under the constraint of a fixed bitrate.

Depending on the structure and philosophy of the coder design, two common approaches to this optimization problem exist: *Bit Allocation* and *Noise Allocation.*

3.2.1 Bit Allocation

The concept of bit allocation is mainly used in the context of simpler coding schemes which do not make use of entropy coding. The idea is to progressively allocate additional bits to groups of spectral coefficients (coder bands) in a way that the perceptual quality of the decoded signal becomes increasingly better after each new allocation step. This process is continued until all available bits are spent. Figure 13 illustrates this process in a flow chart notation.

- In a first step, an initial SNR is set for each band which might be e.g. no allocation of bits for any band (i.e. SNR=0dB).
- From its current SNR setting, the number of bits is calculated for each band. This is typically done in a straight-forward way for uniform quantizers by assuming an increase of 6 dB in SNR for each additional allocated bit per spectral coefficient.
- If the total number of used bits is sufficiently smaller than the number of available bits, another allocation step is performed. Otherwise, the bit allocation procedure terminates and the quantization module is invoked based on the recent bit allocation settings.
- The encoding strategy of the bit allocation routine is determined by the way in which a new allocation step will adjust the SNR settings. Typically, a higher SNR value will be set for bands with the highest deficiency with respect to the perceptual requirements (i.e. the worst NMR value). The "secret" of building a good bit allocation routine is largely related to the selection of good criteria for determining the best candidates for quantization refinement. A collection of conceivable criteria is described in [19].

In general, simple incremental bit allocation algorithms (e.g. based on gradient search methods) may not lead to an optimum solution in a global sense but may terminate at a local optimum. Consequently, more sophisticated methods can be used to avoid this type of suboptimal behavior.

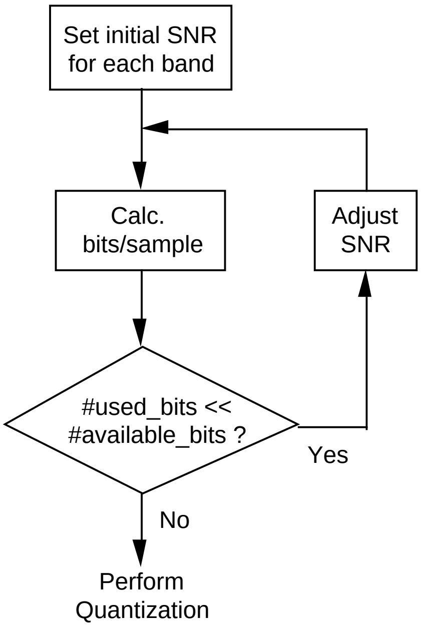
Set initial SNR for each band
Calc. bits/sample
Adjust SNR
#used_bits << #available_bits ?
Yes
No
Perform Quantization

Figure 13: Encoding by bit allocation.

3.2.2 Noise Allocation

As can be seen from the preceding description, the paradigm of bit allocation crucially depends on the availability of an explicit translation between resulting SNR and the number of bits required. While this translation rule (e.g. “6 dB per bit”) is quite obvious in the case of block companding, prediction of the actual bit demand is difficult for coders using entropy coding. For this reason, the encoding strategy of such coders is usually based on the concept of *noise allocation*.

The idea is to run an iterative process which only controls the two relevant coder performance parameters in an explicit way, namely

- the amount of coding noise injected in each band and
- the overall number of bits used for coding the frame.

In contrast to the bit allocation concept, the individual bit consumption of each coder band will not be controlled but will result implicitly from the process.

Figure 14 shows a widely known encoding concept based on noise allocation which is described in the informative annex parts of the MPEG-1 (for Layer III) [1] and MPEG-2 AAC [3] standards. The scheme consists of two nested loops corresponding to a simultaneous optimization in bitrate and perceptual quality:

- In a first step, the quantization precision for all bands is set to an equal value corresponding to a white quantization noise profile.
- Subsequently, a global quantization mechanism is invoked to reduce the general coding precision such that the coding solution stays within the number of available bits. This *rate loop* includes the quantization of all spectral coefficients according to the current quantizer settings, noiseless encoding of the quantized coefficients and calculation of the overall number of bits required for representing the coding solution.
- The generated coding solution is then tested for acceptable perceptual quality (e.g. whether all bands have an associated NMR value below 0 dB). If the criterion is met the algorithm terminates, having produced a solution that both satisfies the perceptual and the rate requirements.
- If the solution is not considered satisfactory in a perceptual sense, a further test is carried out in order to check whether such a solution can be reached at all. If this is not the case (e.g. if all bands have associated NMR values above 0 dB indicating heavy coder overload), the algorithm terminates with a solution which at least meets the rate constraint and thus can be used for transmission.
- Otherwise the iterative improvement of the solution proceeds by refining the quantization resolution for bands with a deficient quantization. Again, this is one of the places in the algorithm were a good strategy will determine the final audio quality achieved. Since refinement of quantization generally leads to an increase in bit demand, processing of this modified solution continues with the rate loop procedure.

As a result of the process, the effective quantization precision (or, equivalently, the amount of amplification before quantization by a fixed quantizer) is delivered for each band as part of the transmitted side information.

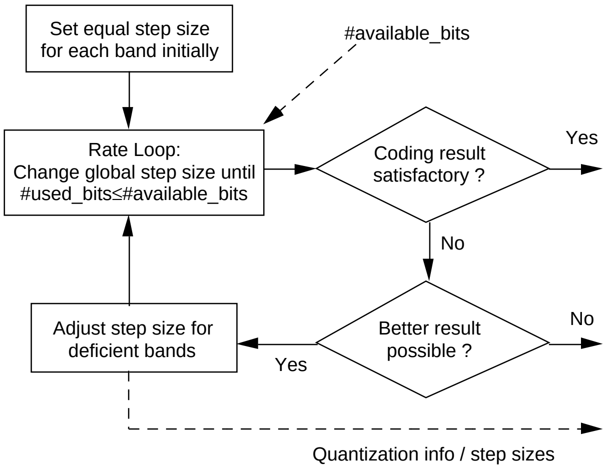
Set equal step size for each band initially
#available_bits
Rate Loop: Change global step size until #used_bits≤#available_bits
Coding result satisfactory ?
Yes
No
Better result possible ?
No
Yes
Adjust step size for deficient bands
Quantization info / step sizes

Figure 14: Encoding by noise allocation.

While the above algorithm consists of two concurrent optimization processes, there are alternative approaches which aim at disentangling both parts and in this way minimize interactions between them. Such a process may consist of the following steps:

- First, the quantization precision for all bands is determined such that the coding distortion is just below the masking threshold (threshold simulator operation). This may involve an iterative search or an a-priori estimation for each band.
- An evaluation of this solution (i.e. counting of the required number of bits) reveals whether coding at the transparent quality level requires more or less than the available number of bits.
- Subsequently, a rate loop will correct the basic quantizer settings in order to meet the target data rate. If less bits were required than available, then additional safety margin can be built up by *overcoding* of the spectral data. Conversely, if the available target rate is exceeded, more noise needs to be allocated (coarser quantization) to save bits. The delivered quality will depend significantly on the choice of an appropriate correction strategy which reduces the bit demand while minimizing the perceptual impact (degradation) of the quantization distortion.

Comparing the two approaches described above, the second approach is very attractive for reasons of its structural and conceptual simplicity. On the other hand, a coding solution with the desired target rate is available only after a complete run through the algorithm including the rate loop whereas the first algorithm can always be terminated after each rate loop and is thus particularly attractive for real-time applications.

3.2.3 Analysis by Synthesis

In order to appropriately control the resulting SNR in each coder band, it is necessary to determine the coding distortion based on a specific setting (i.e. step size) of the quantizer. While this can be done on the basis of expectation values (e.g. N=Q^2/12), many of the more sophisticated systems using entropy coding evaluate the coding distortion by means of an *Analysis by Synthesis* (AbS) approach. This concept is well-known from speech coding (e.g. CELP coders employ AbS) and allows an assessment of the quality of a particular coding solution by synthesizing the coding result and evaluating its properties. In practice, AbS control of audio coder distortion means that the spectral values are quantized / reconstructed and the energy of the actual quantization error is measured by comparison with the unquantized values. While this obviously increases the computational complexity of the algorithm, it will allow precise control over the actual quantization distortion even at very low SNR values when simple estimation of quantization distortion will suffer from large estimation errors. Note that, although unusual, assessment of coding distortion via AbS could also be employed in coders working according to the bit allocation paradigm.

3.3 Bit Reservoir (Constrained Variable Rate Coding)

As was discussed before, coding at a constant perceptual quality generally leads to a *variable rate coding* scheme. Conversely, coding at a *constant bitrate* will usually result in a time dependent coding quality depending on how demanding a particular segment of the input signal is for the coder.

Both concepts can be combined advantageously by using a "constrained variable rate coding" approach which is designed to approach a constant output quality over time. If a frame is easy to code then not all bits are used, but some spare bits are put into a *bit reservoir*. If a frame needs more than the average amount of bits this extra bit allocation is taken from the bit reservoir. The maximum accumulated deviation of the bit demand from the average number of bits in a frame is constrained to a certain value and is referred to as the *size of the bit reservoir*.

This strategy allows local variations in bitrate and helps the coder during times of peak bit demands, e.g. for coding of transient events, while still maintaining a constant average bitrate as required for applications with a constant rate transmission channel. Figure 15 shows an example for the deviation of bit consumption from the average bit rate for an example coder. Note that, as the size of the bit reservoir grows, the coder performance will be able to approach the performance of a constant quality / variable rate coder.

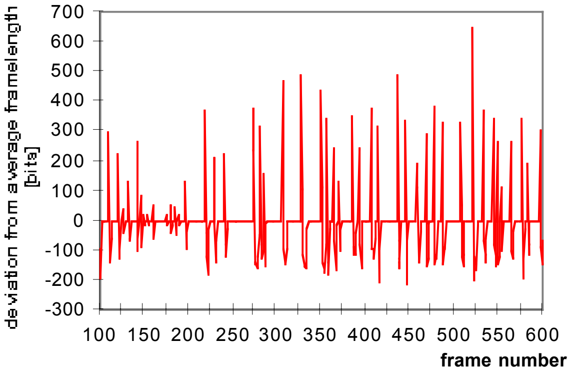

Figure 15: Use of bit reservoir (deviation from average bitrate).

The bit reservoir technique is a well-known part of most coders with high frequency resolution, such as MPEG-1/2 Layer III or MPEG-2 AAC.

Since use of a bit reservoir is equivalent to a local variation in bit rate, the size of the decoder input buffer must be adapted to the maximum local bit rate (i.e. the maximum number of bits which can be allocated for a single frame per channel). The decoder has to wait at least until this input buffer is read before audio output can be started. Consequently, use of the bit reservoir will cause an additional codec delay which grows in proportion to the bit reservoir size [23].

4. CONCLUSIONS

This tutorial overview covered two aspects of perceptual audio coding. Firstly, the general issue of the temporal masking problem in perceptual audio coding was discussed together with the so-called Temporal Noise Shaping (TNS) technology. Secondly, fundamental issues of quantization and coding in a perceptual audio coder were discussed and the most important technical approaches were contrasted. Finally, a discussion of encoding strategies concluded the paper.

Clearly, the flexibility of today's coding schemes will continue to provide the basis for further improvements in coding performance for the foreseeable future.

ACKNOWLEDGEMENTS

The author wishes to express sincere thanks to his colleague Martin Dietz for a number of inspiring suggestions regarding this publication. Thanks also to everyone who helped in proofreading the paper.

REFERENCES

[1] ISO/IEC JTC1/SC29/WG11 (MPEG), International Standard ISO/IEC 11172-3 "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s"

[2] ISO/IEC JTC1/SC29/WG11 (MPEG), International Standard ISO/IEC 13818-3 "Generic Coding of Moving Pictures and Associated Audio: Audio"

[3] ISO/IEC JTC1/SC29/WG11 (MPEG), International Standard ISO/IEC 13818-7 "Generic Coding of Moving Pictures and Associated Audio: Advanced Audio Coding"

[4] ISO/IEC JTC1/SC29/WG11 (MPEG), International Standard ISO/IEC IS 14496-3: "Coding of Audio-Visual Objects: Audio"

[5] ISO/IEC JTC1/SC29/WG11 (MPEG), Committee Draft ISO/IEC 14496-3 Amd 1: "Coding of Audio-Visual Objects: Audio"

[6] K. Brandenburg, G. Stoll, Y.F. Dehéry, J.D. Johnston, L.v.d. Kerkhof, E.F. Schroeder: "The ISO/MPEG-Audio Codec: A Generic Standard for Coding of High Quality Digital Audio", 92nd AES Convention, Wien 1992, Preprint 3336

[7] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, Oikawa: "MPEG-2 Advanced Audio Coding", 101st AES Convention, Los Angeles 1996

[8] K. Tsutsui, H. Suzuki, O. Shimoyoshi, M. Sonohara, K. Akagiri, R. Heddle: "ATRAC: Adaptive Transform Acoustic Coding for MiniDisc", 93rd AES Convention, San Francisco 1992, Preprint 3456

[9] Mark Davis: "The AC-3 Multichannel Coder", 95th AES Convention, New York 1993, Preprint 3774

[10] J. D. Johnston, K. Brandenburg: "Wideband Coding Perceptual Considerations for Speech and Music", in S. Furui and M. M. Sondhi, editors: "Advances in Speech Signal Processing", Marcel Dekker, New York, 1992

[11] J. Herre, J. D. Johnston, "Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)", 101st AES convention, Los Angeles 1996, Preprint 4384

[12] J. Herre, J. D. Johnston: "Exploiting Both Time and Frequency Structure in a System that Uses an Analysis / Synthesis Filterbank with High Frequency Resolution", 103rd AES Convention, New York 1997, Preprint 4519

[13] J. Herre, J. Johnston: "A Continuously Signal-Adaptive Filterbank for High-Quality Perceptual Audio Coding", IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk 1997

[14] N. Jayant, P. Noll: "Digital Coding of Waveforms", Englewood Cliffs, NJ, Prentice-Hall, 1984

[15] Schuyler Quackenbush: "Noiseless Coding of Quantized Spectral Components in MPEG-2 Advanced Audio Coding", IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk 1997

[16] S.-H. Park, Y.-B. Kim, S.-W. Kim, and Y.-S. Seo: Multi-Layer Bit-Sliced Bit-Rate Scalable Audio Coding", 103rd AES Convention, New York 1997, Preprint 4520

[17] N. Iwakami, T. Moriya: "Transform domain weighted interleave vector quantization (TwinVQ)", 101st AES Convention, Los Angeles 1996, Preprint 4377

[18] J. Herre, E. Allamanche, K. Brandenburg, M. Dietz, B. Teichmann, B. Grill, A. Jin, T. Moriya, N. Iwakami, T. Norimatsu, M. Tsushima, T. Ishikawa: "The Integrated Filterbank Based Scalable MPEG-4 Audio Coder", 105th AES Convention, San Francisco 1998, Preprint 4810

[19] Stephen Voran: "Perception-Based Bit-Allocation Algorithms For Audio Coding", IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk 1997

[20] E. Zwicker, E. Terhardt: "Analytical Expressions for Critical Bandrate and Critical Bandwidth as a Function of Frequency," J. Acoust. Soc. of America 68 (1980), pp. 1523-1525

[21] B. Moore: "Characterisation of Simultaneous, Forward and Backward Masking", Proc. of the 12th International AES Conference on The Perception of Reproduced Sound, Kopenhagen, pp. 22-33, 1993

[22] J. Princen, A. Johnson, A. Bradley: "Subband / Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation", IEEE ICASSP 1987, pp. 2161 - 2164

[23] E. Allamanche, R. Geiger, J. Herre, T. Sporer: "MPEG-4 Low Delay Audio Coding based on the AAC Codec", 106th AES Convention, Munich 1999, Preprint 4929

[24] J. D. Johnston: "Estimation of Perceptual Entropy Using Noise Masking Criteria", IEEE ICASSP 1988, pp. 2524 - 2527

[25] J. Herre, E. Eberlein, K. Brandenburg: "A Real Time Perceptual Threshold Simulator", IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk 1991

[26] B. Edler: "Codierung von Audiosignalen mit überlappender Transformation und adaptiven Fensterfunktionen", Frequenz, Vol. 43, pp. 252-256, 1989 (in German)