Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)

4384 (N-3)

Jürgen Herre
Lucent Technologies Bell Laboratories
Murray Hill, NJ 07974, USA

James D. Johnston
AT&T Laboratories
Murray Hill, NJ 07974, USA

Presented at
the 101st Convention
1996 November 8–11
Los Angeles, California

AES

This preprint has been reproduced from the author's advance manuscript, without editing, corrections or consideration by the Review Board. The AES takes no responsibility for the contents.

Additional preprints may be obtained by sending request and remittance to the Audio Engineering Society, 60 East 42nd St., New York, New York 10165-2520, USA.

AN AUDIO ENGINEERING SOCIETY PREPRINT

Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)

Jürgen Herre* and James D. Johnston**
* Lucent Technologies Bell Laboratories
** AT&T Laboratories

Abstract

Inappropriate temporal spread of quantization noise is known to be the reason for the so-called "pre-echo" artifacts in perceptual audio coding. These artifacts occur when a transient signal is being coded in a spectral representation because the quantization noise is spread out over the entire window length of the filterbank and in this way is not masked by the signal. A novel approach to the pre-echo problem is presented allowing to shape the quantization noise in time. The technique is shown to work efficiently also in the case of "pitched" signals (e.g. speech) where traditional block switching schemes do not offer an efficient solution.

Introduction

During the last years so-called perceptual audio coders have been developed enabling the transmission and storage of high quality audio signals at bit rates of about 1/12 and less of the bit rate used on the Compact Disc medium (CD) [Bra91] [Joh92]. Respective standards have been established under ISO/MPEG [MPEG1] [MPEG2] or are currently in progress [MP2NBC] [MPEG4].

Despite the advanced state of today's perceptual audio coders the handling of transient input signals still presents a major challenge to these schemes. This is mainly due to the problem of maintaining the masking effect in the reproduced audio signal. In particular, coding is difficult because of the temporal mismatch between masking threshold and quantization noise.

In this paper we will discuss the underlying problem of temporal masking in the context of perceptual coders. Traditional approaches to this problem will be reviewed and discussed in terms of benefits and drawbacks. Finally, a novel method for controlling the temporal spread of the quantization noise is presented. This technique for "temporal noise shaping" (TNS) is discussed in its basic concepts, relationship to known techniques and implementation in a perceptual audio coder. First results on the coding quality are reported.

The Temporal Masking Problem in Perceptual Audio Coding

Generic coder structure

Generally, the structure of a perceptual audio coder for monophonic audio signals can be described as follows (see figure 1):

- The input samples are converted into a subsampled spectral representation using various types of filterbanks and transforms as e.g. the modified discrete cosine transform (MDCT, [Pri87]), polyphase filterbanks [Rot83] or hybrid structures [Bra90] [Bra92].
- Using a perceptual model the signal's time-dependent masked 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 quantized and coded according to the precision corresponding to the masked threshold estimate. In this way, the quantization noise is hidden (masked) by the respective transmitted signal and is thus not perceptible after decoding.
- 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 processing used in the encoder is reversed in the decoder (see figure 2):

- 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.

Using this generic coder structure it is possible to efficiently exploit the irrelevancy contained in each signal due to the limitations of the human auditory system. Specifically, the effect of *simultaneous masking* [Yos77][Moo89] can be utilized by shaping the spectrum of the quantization noise according to the shape of the signal's masking threshold (known as the *spectral noise shaping problem*) [Sch79] [Joh96]. In this way, the noise is "hidden" under the coded signal and perceptually transparent quality can be achieved even at very low data rates like 64 kbit/s for a monophonic signal.

Pre-echoes: The temporal masking problem

While the classic perceptual coder is primarily designed to exploit the perceptual effect of *simultaneous masking* the masking phenomenon also has a temporal aspect: Noise is masked a short time prior to and some time after the presentation of a masking signal (*pre-masking* and *post-masking* phenomenon) [Yos77][Moo89]. Figure 3 shows a principle plot of masking over time for a switched signal (masker). Post-masking is observed for a much longer period of time than pre-masking (in the order of 10-50 ms instead of 0,5-2 ms, depending on the level and duration of the masker).

Thus, the temporal aspect of masking leads to an additional requirement for a perceptual coding scheme: In order to achieve perceptually transparent coding quality the quantization noise also must not exceed the time-dependent masked 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 point MDCT) this means that the quantization noise may be spread out over a period of more than 40 milliseconds. This will lead to problems when the signal to be coded contains strong signal components only in parts the analysis filterbank window, i. e. for transient signals. In particular, quantization noise is spread out *before* the onsets of the signal and in extreme cases may even exceed the original signal components in level during certain time intervals. As an example, figure 4 shows a plot of a coded and decoded castanets signal. As can be seen, the noise components are spread out a certain time before the “attack” of the original signal. Such a constellation is traditionally known as a “pre-echo phenomenon” [Joh92b].

Figure 4 shows an example of this phenomenon that has been generated by removing the pre-echo control facilities from a coder and plotting the energy of the signal and the coding noise over time. Obviously, the noise level increases a considerable amount of time before the transient signal onset since it is spread across the entire analysis block.

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 an appropriate temporal characteristics 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 bit-rates for transient signals like castanets, glockenspiel, triangle etc.

Traditional approaches to the temporal masking problem

A set of techniques has been proposed in order to avoid pre-echo artifacts in the encoded / decoded signal:

Pre-echo control and bit reservoir

One way is to increase the coding precision for the spectral coefficients of the filterbank window that first covers the transient signal portion (so-called "pre-echo control", [MPEG1]). Since this considerably increases the amount of necessary bits for the coding of such frames this method cannot be applied in a constant bit rate coder. To a certain degree, local variations in bit rate demand can be accounted for by using a bit reservoir ([Bra87], [MPEG1]). This technique permits to handle peak demands in bit rate using bits that have been set aside during the coding of earlier frames while the average bit rate still remains constant. In practice, however, the size of the bit reservoir must to be unrealistically large in order to avoid artifacts when coding input signals of very transient nature.

Adaptive window switching

A different strategy used in many perceptual audio coders is adaptive window switching as introduced by Edler [Edl89]. This technique adapts the size of the filterbank windows to the characteristics of the input signal. While stationary signal parts will be coded using a long window length, short windows are used to code the transient part of the signal. In this way, the peak bit demand can be reduced considerably because the region for which a high coding precision is required is constrained in time.

Figure 5 shows an example for a common window sequence using adaptive window switching. Between long and short windows intermediate window shapes ("start block", "stop block") .have to be selected to ensure proper signal reconstruction. The example shown below uses long and short window sizes of 2048 samples and 256 samples, respectively.

One major disadvantage of the adaptive window switching technique is that it introduces additional complexity into the coder and complicates its structure. Since different window sizes require different interpretations and normalizations of the psychoacoustic model, as well as different frequency band and noiseless coding structures, window switching complicates the coder structure noticeably. In addition, the need to make switching decisions in the presence of an overlap-add structure filterbank (i.e. MDCT, ELT [Mal92b] ...) requires additional buffering and delay in the encoder, resulting in more end-to-end delay. Finally, while the "long" and "short" windows have good time / frequency localization properties, the "start" and "stop" windows do not and introduce even more coding inefficiency.

Furthermore, a limitation of adaptive window switching is given by its latency and repetition time: Using the window types of the previous example, the fastest possible turn-around cycle between two short block sequences requires at least three blocks ("short"→ "stop"→ "start"→ "short", ca. 30 - 60 ms for typical block sizes of 512 -

1024 samples) which might be much too long for certain types of input signals. As an example, figure 6 shows a waveform plot of a speech excerpt ("German male speech"). Due to the mechanism of speech production, the voiced parts consist of an impulse-like excitation signal filtered by an acoustic tube resulting in a continuous stream of transient events (we will refer to such signals as "pitched" signals). For speech, the period between subsequent excitation pulses (ca. 8 ms in our example) is definitely below the turn-around time of a window switching scheme. Consequently, temporal spread of quantization noise can only be avoided by permanently or frequently selecting the short window size, which usually leads to a decrease in the coder's source-coding efficiency. This is due to the loss in filterbank gain (diagonalization) and the need for more side information, for all parts of the signal that do not require good temporal control of the quantization noise [Joh96].

Gain modification (gain control)

A third way to avoid the temporal spread of quantization noise is to apply a dynamic gain modification (gain control process) to the signal prior to calculating its spectral decomposition [Vau91] [Lin93].

The principle of this approach is illustrated in figure 7. The dynamics of the input signal is reduced by a gain modification (multiplicative preprocessing) prior to its encoding. In this way, "peaks" in the signal are attenuated prior to encoding. The parameters of the gain modification are transmitted in the bitstream. Using this information the process is reversed on the decoder side.

Effectively, applying a dynamic multiplicative modification of the input signal prior to its spectral decomposition is equivalent to a dynamic modification of the filterbank's analysis window. Depending on the shape of the' gain modification function the frequency response of the analysis filters is altered according to the composite window function.

In order to perform well for many signals, however, it is important that the gain modification processing can be applied independently in different parts of the audio spectrum because

- transient events are often dominant only in parts of the spectrum
- it is undesirable to widen the frequency response of the filterbank's low frequency filter channels because this increases the mismatch to the critical bandwidth

Frequency dependent gain modification can be achieved at the expense of an increase in complexity by using a hybrid filterbank structure. In this way, independent gain processing can be carried out on each subband signal after the first stage of the hybrid filterbank [Aka95].

MPEG-Audio Coder Techniques

In the coder family defined by MPEG-1 and MPEG-2 Audio [MPEG1] [MPEG2], Layers I and II use a low frequency resolution filterbank and do not include special measures for handling pre-echo situations. Layer III employs a high frequency resolution filterbank together with a combination of pre-echo control, bit reservoir and adaptive window switching.

The Temporal Noise Shaping (TNS) Technique

In this chapter we will present a novel concept for addressing the temporal noise shaping problem as described above. First the basic concept is outlined together with the theoretical foundations, then the implementation in a generic perceptual audio coder is described. Finally, the relationship to previously known techniques for signal processing will be discussed.

Basic Concept

Optimum Coding Methods for Transient Signals
Let us first review the mathematical background that leads to an extended formulation of optimal coding methods for transient signals.

Given a real signal *x(t)* it can be shown that the square of its Hilbert envelope, *e(t)*, can be expressed as

$$e(t) = F^{-1}\left\{ \int C(\zeta) \cdot C^{*}(\zeta - f) \; d\zeta \right\} \qquad (1)$$

where *C(f)* is the single sided spectrum of *x(t)* for positive frequencies (see annex for a detailed derivation). In other words, the Hilbert envelope of a signal is directly connected to the autocorrelation function of its spectrum.

Note that this relation is the dual to the well-known formula relating the power spectral density *PSD(f)* of a signal to its autocorrelation function in time domain.

$$PSD(f) = F\left\{ \int x(\tau) \cdot x^{*}(\tau - t) \; d\tau \right\} \qquad (2)$$

Thus, the squared Hilbert envelope of a signal and the power spectral density constitute dual aspects in time and frequency domain.

Assuming a discrete representation of the signals, the following conclusion can be drawn from equation (1): If the Hilbert envelope remains constant for each partial bandpass signal across a range of frequencies then also the autocorrelation between adjacent spectral values will remain constant. This in facts means that the series of spectral coefficients is stationary across frequency and thus predictive coding techniques can efficiently be used to represent this signal using one common set of prediction coefficients.

Figures 8 and 9 illustrate that this property is indeed approximately fulfilled for transient signals. The first figure shows a short excerpt from a transient "castanets" signal of ca. 40 ms duration. This signal has been decomposed into several partial bandpass signals of

500 Hz bandwidth each. Figure 9 shows the Hilbert envelopes for these bandpass signals with center frequencies ranging from 1500 Hz to 4000 Hz (for clarity, all envelopes have been normalized in their maximum amplitude). Obviously, the shapes of all partial envelopes are related very strongly so that a common predictor can be used within this frequency range to efficiently code the signal. Similar observations can be made for speech signals where the effect of the glottal excitation pulses is present across the entire frequency range due to the nature of the human speech production mechanism (see figure 6).

An alternative way of gaining an intuitive understanding of the spectral predictability property of transient signals is to take a look at the following dualities (table 1). Let us first consider a sinusoidal input signal with a flat temporal envelope. The corresponding frequency domain representation consists of a dirac impulse, i.e. a maximally non-flat spectral shape. Such a signal is known to be easily represented by either directly coding its spectral data ("transform coding") or by applying linear predictive coding (LPC) to the time signal [Jay84]. Now let us consider the opposite case where the input signal is a dirac impulse (in time) corresponding to a "flat" power spectrum (the phase spectrum rotates according to the temporal position of the impulse). Obviously, this signal constitutes a worst-case signal for the above mentioned traditional methods like transform coding (coding of spectral data) or LPC (predictive coding of time domain data). The optimum coding techniques may be derived, however, by swapping time and frequency domain which leads to either direct coding of time domain data or *predictive coding of spectral data* (see above). The coding technique presented in this paper relies on the latter of both alternatives.

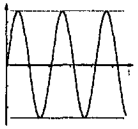
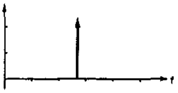
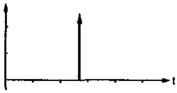
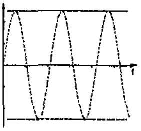
Input Signal		**Optimum Coding**	
Time Domain	Freq. Domain	Direct Coding	Predictive Coding
		Coding of spectral data	Prediction in time domain
		Coding of time domain data	*Prediction in* *frequency domain*

Table 1: Optimum coding methods for extreme input signal characteristics

It is important not to confuse predictive coding of spectral coefficients over frequency with the well-known dual concept of Subband-ADPCM [Ata81]. In the latter case prediction is carried out separately for each spectral coefficient over time. *While such a prediction over time increases spectral resolution, prediction over frequency enhances temporal resolution.*

Temporal Noise Shaping via Prediction in the Spectral Domain
In the previous paragraph we used the dualities between time and frequency domain to identify an efficient coding method for transient signals: Predictive coding techniques can advantageously be applied to the signal's spectral data over frequency very much in the same way they are employed in traditional Differential Pulse Code Modulation (DPCM) coders for the coding of stationary signals in time.

Once the dualities between time and frequency domain are established we can make use of the well-understood concept of DPCM coding of time signals [Jay84] to determine the properties of à coder applying DPCM over frequency for the coding of spectral coefficient data. Since predictive coding is applied to spectral domain data, the relations known for classic prediction are valid with time and frequency domain swapped:

- Coding gain
  As the power spectral density (PSD) and the squared Hilbert envelope are dual to each other, a reduction of residual signal energy ("prediction gain") is achieved depending on the "squared-envelope flatness measure" of the signal (as opposed to the "spectral flatness measure" [Jay84]). Thus, the potential coding gain of the DPCM coder increases with the "non-flatness" (transient character) of the squared time envelope of the signal.

- Possible prediction schemes
  In the case of transient signals, both the traditional closed-loop and open-loop prediction structures [Jay84] can be employed to provide a reduction of residual energy (coding gain). Figures 10 a) and b) show encoder diagrams for both schemes. Depending on the selection of the prediction scheme, however, different advantages will be achieved in the overall system behavior:

- Closed-loop prediction scheme
  If a closed-loop prediction scheme is used the error energy in the decoded signal will decrease according to the achieved prediction gain. Since the error introduced into the spectral coefficient data has a "flat" PSD, also the *temporal* shape of the quantization noise signal remains "flat" in the decoded signal after the synthesis filterbank. In other words, the error signal energy is distributed uniformly over time just like in the case without DPCM. This is, however, of limited use in the psychoacoustic sense because it *does not prevent temporal unmasking of the quantization noise* during silent portions of the coding window (see "temporal noise shaping problem" discussed above).

- Open-loop prediction scheme
  If an open-loop quantization scheme is used instead, the total error energy in the

decoded spectral coefficient data will remain the same, i.e. there is no gain in terms of overall error energy. In contrast to the closed-loop approach, the temporal shape of the quantization error signal will, however, appear as shaped in time at the output of the decoder (since the DPCM processing has been applied to *spectral* coefficients, the quantization noise in the signal after the inverse filterbank will be shaped in *time*), putting the quantization noise under the actual signal. In this way, problems of temporal masking, either in transient or pitched signals, are avoided. This type of predictive coding of spectral data will therefore be referred to as the "Temporal Noise Shaping" (TNS) algorithm.

An illustration of the general principle is given in figures 11 and 12. In this example, Temporal Noise Shaping is applied to a castanets attack above a frequency of 4 kHz using a 1024 point DCT filterbank.

- Figure 11 shows the input signal (left top) and the resulting frequency response of the synthesis filter (left bottom) as computed by the LPC calculation. Please note that the (normalized) frequency coordinates in this plot correspond to the time coordinates due to the time / frequency duality. Obviously, the LPC calculation leads to a "source model" of the input signal reflecting its average envelope shape.
- A comparison between the original spectral coefficients (right top plot) and residual signal after prediction (right bottom plot) shows a consistent reduction of residual energy corresponding to an overall prediction gain of ca. 12 dB.
- Next, coding noise was injected into the (residual) spectral coefficients so that an SNR of ca. 13 dB resulted in each of the coder bands (width ca. 1/2 Bark). The resulting error signals are shown in figure 12 both for the case with and without Temporal Noise Shaping. As expected from earlier 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.

Figure 13 illustrates the application of the Temporal Noise Shaping principle to a speech signal ("German Male Speech") using a 1024 point MDCT filterbank. 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.

Using Temporal Noise Shaping in a Perceptual Coder

Temporal Noise Shaping can easily be incorporated into a standard perceptual coder by making the following extensions to it (see figures 1 and 14):

- The input signal is decomposed into spectral coefficients by a high-resolution filterbank / transform. Based on the computed spectral coefficients a standard LPC calculation is carried out (e.g. forming of the autocorrelation matrix and using a

Levinson-Durbin recursion) on the spectral coefficients belonging to the target frequency range (e.g. 1.5 - 20 kHz). This is done for the highest permitted order of the noise shaping filter (e.g. 20). If the calculated prediction gain exceeds a certain threshold Temporal Noise Shaping is activated.

- The order of the used noise shaping filter for the current block is then determined by subsequently removing all reflection coefficients with a small enough absolute value from the "tail" of the reflection coefficient array. In this way, the remaining filter order typically ranges between 4 - 12 for a speech signal.
- In this case, the encoding of the spectral values is done by replacing the standard PCM quantizer by a DPCM scheme operating on the filterbank outputs in frequency. In practice, this can easily be accomplished by sliding the prediction filter (as calculated in the LPC calculation) across the set of spectral coefficients to which Temporal Noise Shaping should be applied, so that the original coefficients are replaced by the residual signal and passed on to the standard quantizer. This is symbolized in figure 14 by a "rotating switch" circuitry. Both sliding in the order of increasing and decreasing frequency is possible.
- The side information transmitted to the decoder is extended by a flag indicating the use of TNS and, if used, information about the target frequency range and the filter employed for encoding. The filter data may be represented e.g. as quantized PARCOR coefficients [Par87].

In the decoder, the following processing steps are added for each channel to use Temporal Noise Shaping (see figures 2 and 15):

- Decoding of TNS related side information (flag indicating the use of TNS and, if used, information about the target frequency range and the filter employed for encoding).
- If TNS is active the inverse quantization of the spectral coefficients is carried as DPCM decoding. In practice, this is accomplished by sliding the inverse prediction filter across the set of residual spectral coefficients to which Temporal Noise Shaping should be applied, so that the transmitted residual signal is replaced by the decoded signal and are passed on to the inverse filterbank. Again, this is symbolized by a "rotating switch" circuitry in figure 15. Both sliding in the order of increasing and decreasing frequency is possible in accordance with the direction used in the encoder.

Temporal Noise Shaping and Lapped Orthogonal Transforms

In the previous chapters the discussion of Temporal Noise Shaping Techniques was based on the notion of Fourier transform (and Discrete Fourier Transform DFT in the case of discrete spectral coefficients). In practice, the MDCT is preferred over the FFT and/or DCT in a modern transform coder for the reasons that it is both critically sampled and at the same time maintains frequency meaning on both analysis and synthesis, including across block boundaries [Mal92].

It can be shown that the Temporal Noise Shaping Technique provides a straight forward temporal noise shaping effect also for the known classic orthogonal block transforms like Discrete Fourier Transform (DFT) or Discrete Cosine or Sine Transform (DCT, DST). 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 [Pri87]) 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 used synthesis window. Thus it is advantageous to choose a filterbank window that exhibits only a small overlap between subsequent blocks so that the temporal aliasing effect is minimized. An appropriate algorithmic strategy in the encoder can e.g. adaptively select a window with a low degree of overlap for critical signals of very transient character while using a "wider" window type for stationary signals providing a better frequency selectivity. Figure 16 shows three types of valid TDAC windows with a varying degree of overlap.

Relationship to Gain Modification Technique

Comparing the Temporal Noise Shaping method with the gain modification technique described earlier, it becomes obvious that both are related by the time/frequency duality:

- The gain modification technique carries out a multiplicative modification of the input samples.
- Temporal Noise Shaping performs a convolution of the spectral coefficient data.

In effect, the TNS technique is equivalent to applying an adaptive time domain window determined by the LPC calculation in the frequency domain (since the prediction filter is generally not linear-phase the corresponding time window is complex).

Contrary to the "gain modification" approach, however, there is no need to employ a hybrid filterbank to allow frequency dependent processing of the signal. The application range of the process is also not restricted to the fixed frequency bands defined by the first stage of the hybrid filterbank.

Adaptive vs. Non-adaptive Filterbanks

A frequently discussed topic in perceptual coding is the question about the optimal filterbank. Obviously, the answer depends on the input signal and can be stated for two extreme types of signals as follows [Joh96]:

- For stationary or pseudo-stationary signals with many frequency components (harmonics) like Harpsichord, it is essential to have a sufficiently large transform size to resolve the lines in the signal spectrum in order to extract the redundancy. Protection from time domain artifacts does not play a major role on average for these signals. Thus, a high-resolution uniform filterbank is a good choice in such cases.
- For transient signal types like e.g. castanets, the emphasis of the coding process is on the removal of irrelevance by optimally exploiting the masking properties of the human auditory system. Since frequency resolution in the high frequency range is not available in such cases, the optimum choice for such cases is a critical band filter structure.
  In practice, because of the ease of implementation and computational complexity constraints, many coders use a uniform lower-resolution filterbank instead of the critical band structure for coding of transient signal parts.

Obviously, the optimum filterbank for any input signal ranges somewhere in the continuum between these extreme cases. In a standard perceptual coder using window switching, however, no soft transition between both cases is possible but both cases are handled by "hard" switching between high and low frequency resolution or between a uniform and a non-uniform filterbank [Sin96].

This limitation is overcome by using Temporal Noise Shaping in the following way:

- As seen in the last section, applying TNS is equivalent to adaptively adjusting the time window of the filterbank which in turn influences the frequency response of the filterbank's prototype filter.
- For transient signals, the equivalent window shape is modified heavily and the filter curve gets "widened". In this case, the filterbank's increased temporal resolution 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.
- Thus, 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" approach.

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 mentioned above for the optimum filterbank for a given input signal.

Properties of Temporal Noise Shaping

The key features of the Temporal Noise Shaping approach can be summarized as follows:

- 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.
- The method reduces the peak bit demand of the coder for transient signal segments by exploiting irrelevancy. As a side effect, the coder can stay longer in the preferred "long block" mode so that use of the critical "short block" mode can be minimized.
- 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 the 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.

Results

Preliminary tests of the Temporal Noise Shaping algorithm suggest that this technique allows for a substantial reduction in required overcoding for transient or pitched signals. In comparison, even with block switching, overcoding of the signal by 10-50dB is often required in an attack, and overcoding by 9-12dB is required for noise-like parts of a pitched signal.

The performance of the TNS technique has been evaluated in the course of the standardization process of the ISO/MPEG2-Audio NBC ("non backward compatible coding") coder [MP890]. The core experiment compared the sound quality of a standard "Reference Model 3" (RM3) coder with an enhanced version using TNS at a bitrate of 64 kbit/s for a monophonic signal. Figure 17 shows the Mean Opinion Score (MOS) values for the test set comprising 6 extremely critical test sequences (Harpsichord, Castanets, German Male Speech, Bagpipes, Glockenspiel, Pitch Pipe). In previous experiments, the "German Male Speech" recording had been found to be particularly demanding because of the combination of both low pitch frequency and sharply defined pitch pulses such that the quantization noise tends to become unmasked between the pulses. Although the coding performance for many test items was already very high (grade of 4 and above on the five-grade impairment scale) the speech item was graded fairly low (MOS 2.64) due to a "double speak" phenomenon where a "second whispering speaker" (coding noise) was perceived in the decoded signal. The introduction of a first version of TNS into the reference model coder lead to an improvement of ca. 0.9 MOS grades on this item thus addressing the most obvious "weak spot" in the coder. In addition, an improvement of 0.3 was achieved on "Glockenspiel" (MOS 3.85 → 4.15) by improving the sound quality during the transient signal portions. A quality gain for the castanets signal could, however, not be realized in this core experiment because the quality for this item was largely determined by other limiting factors present in the coder's implementation structure. For the rest of the items, no statistically significant change was observed (overlapping confidence intervals) due to the addition of TNS, as it can be expected for signals without distinct temporal fine structure.

Temporal Noise Shaping has been adopted as a part of the NBC coder scheme [Bos96] [MP2NBC].

Conclusions

A novel technique for addressing the temporal masking problem in perceptual audio coding has been presented. The technique is shown to work efficiently also in the case of "pitched" signals (e.g. speech) where traditional block switching schemes do not offer a solution. The performance gain of the Temporal Noise Shaping technique has been verified in the recent development process of the MPEG2-Audio NBC coder. Although the discussion in this paper has been entirely focused on coding of audio signals, the general principle of temporal noise shaping can also be applied to different fields of perceptual coding as e.g. image coding (addressing "edge effects").

Acknowledgments

The authors wish to express their thanks to David Malah of the Technion Israel Institute for Technology and to their colleague Peter Kroon for their valuable assistance regarding predictive coding methods. They would also like to thank Emre Telatar of the Bell Laboratories Mathematics Center for his analysis of Hilbert envelope mathematics which subsequently inspired the development of Temporal Noise Shaping.

Annex: Frequency Domain Description of Hilbert Envelope

Given signal a real signal *x(t)* and its complex envelope (analytic signal) *c(t)*

$$c(t) = x(t) + j \cdot H\{x(t)\}$$

where *H*{ *x(t)* } denotes the Hilbert transform of *x(t)* defined by

$$H\{\ x(t)\ \} = \int \frac{x(\tau)}{\pi(t-\tau)} d\tau$$

If *x(t)* has the Fourier transform *X(f)* then the Fourier transform *C(f)* of the complex envelope is a one-sided spectrum

$$C(f) = \begin{cases} 2X(f) & f > 0 \\ X(f) & f = 0 \\ 0 & f < 0 \end{cases}$$

Now let us define *e(t)* as the square of the corresponding Hilbert envelope (i.e. absolute value of the analytic signal):

$$e(t) = |\ c(t)\ |^2 = c(t) \cdot c^*(t)$$

Then the Fourier transform *E(f)* of the squared envelope can be calculated as

$$E(f) = C(f)\ *\ C^*(f) = \int C(\zeta) \cdot C^*(\zeta - f)\ d\zeta$$

and *e(t)* itself can be expressed as

$$e(t) = F^{-1}\left\{\ \int C(\zeta) \cdot C^*(\zeta - f)\ d\zeta\ \right\}$$

In other words, the signal envelope is directly connected to the autocorrelation in the spectral domain (it is in fact the inverse Fourier transform of the spectral autocorrelation function).

References

[Aka95] ISO/IEC JTC1/SC29/WG11 MPEG input document: K. Akagiri: "Technical Description of Sony Preprocessing", MPEG95/026, Dallas

[Ata81] B. S. Atal, J. R. Remde: "Split-Band APC System for Low Bit Rate Encoding of Speech", Proc. IEEE ICASSP, April 1981

[Bos96] 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

[Bra87] K. Brandenburg: "OCF - A New Coding Algorithm for High Quality Sound Signals", Proc. IEEE ICASSP, 1987

[Bra90] K. Brandenburg, J. D. Johnston: "Second Generation Perceptual Audio Coding: The Hybrid Coder", 88th. AES Convention, Montreux 1990, Preprint 2937

[Bra91] K. Brandenburg, J. Herre, J.D. Johnston, Y. Mahieux, E.F. Schroeder: "ASPEC: Adaptive Spectral Perceptual Entropy Coding of High Quality Music Signals", 90th. AES Convention, Paris 1991, Preprint 3011

[Bra92] K. Brandenburg, E. Eberlein, J. Herre, B. Edler: "Comparison of Filterbanks for High Quality Audio Coding", IEEE ISCAS, San Diego, 1992

[Edl89] B. Edler: "Codierung von Audiosignalen mit überlappender Transformation und adaptiven Fensterfunktionen", Frequenz, Vol. 43, pp. 252-256, 1989

[Jay84] N. Jayant, P. Noll: "Digital Coding of Waveforms", Englewood Cliffs, NJ, Prentice-Hall, 1984

[Joh92] J. D. Johnston, A. J. Ferreira: "Sum-Difference Stereo Transform Coding", IEEE ICASSP 1992, pp.569-571

[Joh92b] 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

[Joh96] J. D. Johnston: "Audio Coding with Filter Banks", pages 287-307 in: "Subband and Wavelet Transforms" by A. N. Akansu and M. J. T. Smith (editors), Kluwer Academic Publishers, Norwell 1996

[Lin93] M. Link: "An Attack Processing of Audio Signals for Optimizing the Temporal Characteristics of a Low Bit-Rate Audio Coding System", 95th AES convention, New York 1993, Preprint 3696

[Mal92] H. S. Malvar: "Signal Processing with Lapped Transforms", Artech House, Norwood, MA, 1992

[Mal92b] H. Malvar: "Extended Lapped Transforms: Properties, Applications, and Fast Algorithms", IEEE Transactions on Signal Processing, Vol. 40, No. 11, November 1992

[Moo89] B. C. J. Moore: "An Introduction to the Psychology of Hearing, Academic Press, London, 1989

[MP2NBC] ISO/IEC JTC1/SC29/WG11 MPEG, Committe Draft ISO 13818-7 "Generic Coding of Moving Pictures and Associated Audio: Audio" (non backwards compatible coding, NBC)

[MPEG1] ISO/IEC JTC1/SC29/WG11 MPEG, International Standard ISO 11172-3 "Coding of moving pictures and associated audio for digital storage media at up to about 1.5Mbit/s"

[MPEG2] ISO/IEC JTC1/SC29/WG11 MPEG, International Standard ISO 13818-3 "Generic Coding of Moving Pictures and Associated Audio: Audio"

[MPEG4] ISO/IEC JTC1/SC29/WG11 MPEG: "MPEG-4 Proposal Package Description (PPD)", document N0998

[MP890] ISO/IEC JTC1/SC29/WG11 MPEG: C. Lueck, M. Ali: "NBC Reference Model 3 monophonic subjective tests: overall results", document M890

[Par87] T. W. Parson: "Voice and Speech Processing", McGraw-Hill Book Company, New York 1987

[Pri87] 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

[Rot83] J.H. Rothweiler: "Polyphase Quadrature· Filters - a new Subband Coding Technique", IEEE ICASSP 1983, Boston, pp. 1280 - 1283

[Sin96] D. Sinha, J. D. Johnston: "Audio Compression at Low Bit Rates Using a Signal Adaptive Switched Filterbank", Proc. IEEE ICASSP, 1996, pp. 1053ff

[Sch79] M. R. Schroeder, B. S. Atal, J. L. Hall: "Optimizing Speech Coders by Exploiting Masking Properties of the Human Ear", Journal of the Acoustic Society of America 66 (1979), pages 1647-1652.

[Vau91] T. Vaupel: "Ein Beitrag zur Transformationscodierung von Audiosignalen unter Verwendung der Methode der 'Time Domain Aliasing Cancellation (TDAC)' und einer Signalkompandierung im Zeitbereich", PhD Thesis, Universität-Gesamthochschule Duisburg, Germany, 1991

[Yos77] W. A. Yost: "Fundamentals of Hearing", Academic Press Inc., New York, Boston London etc. 1977, 1985, 1994

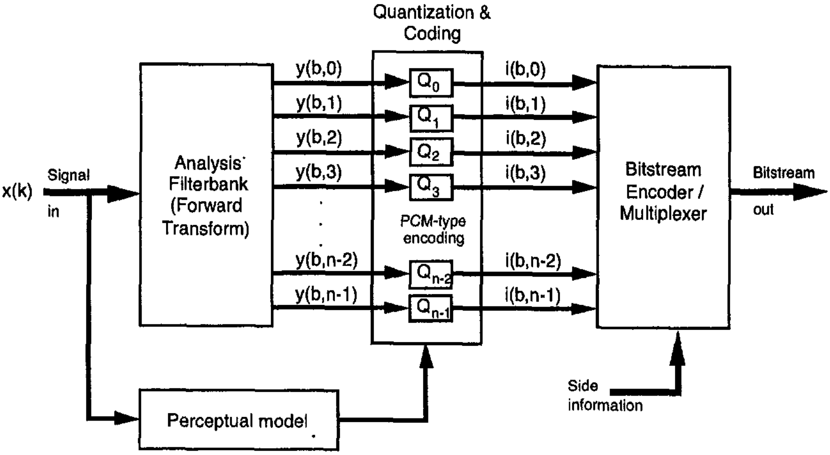


Figure 1: Generic (monophonic) perceptual encoder

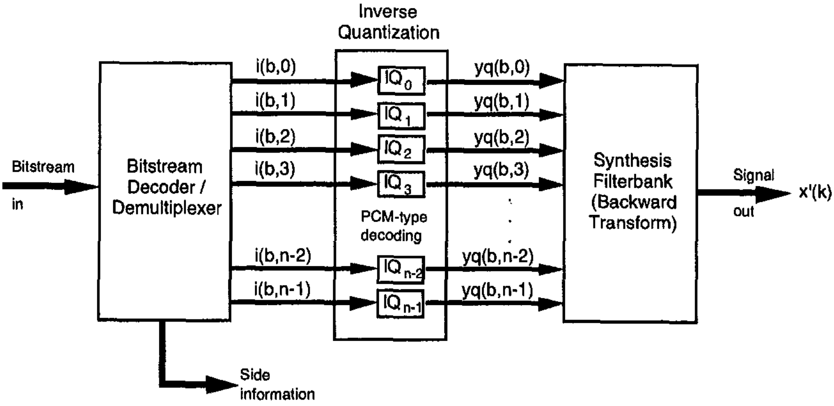


Figure 2: Generic (monophonic) decoder

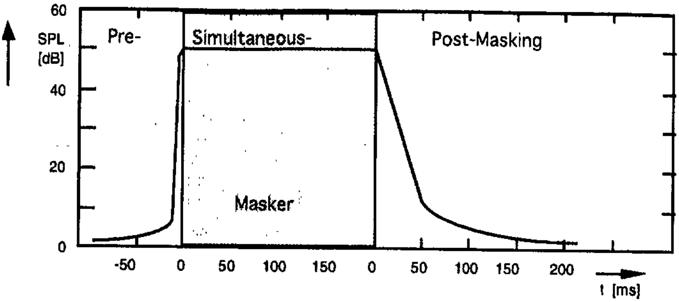


Figure 3: Principle of pre-masking, simultaneous masking and post-masking

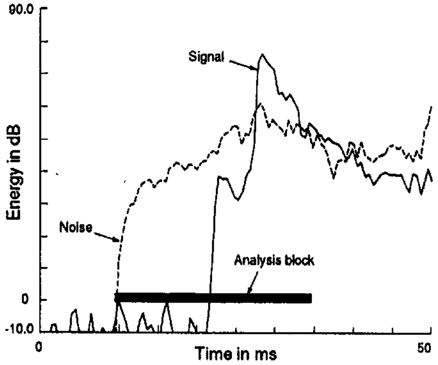


Figure 4: Example of "pre-echo" phenomenon

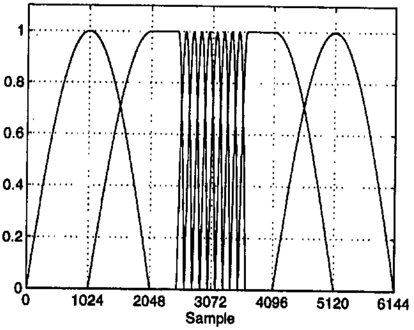


Figure 5: Example for adaptive window switching sequence

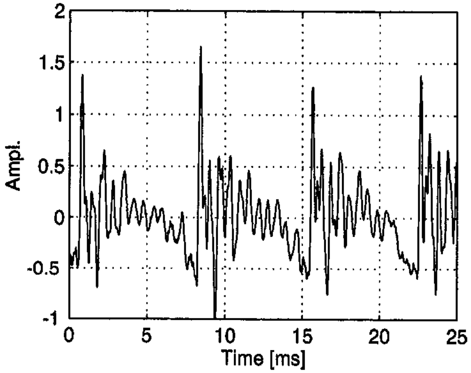


Figure 6: Speech signal ("German Male Speech")

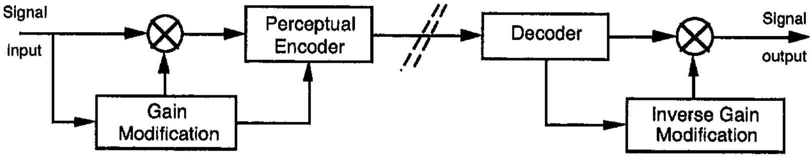


Figure 7: Principle of gain modification technique

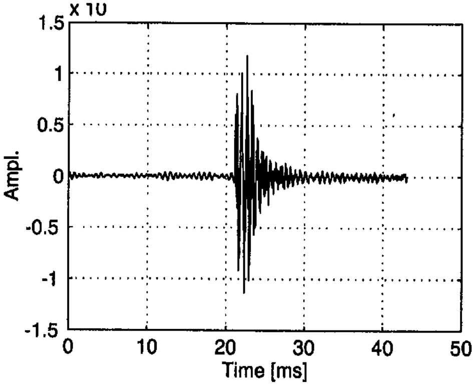


Figure 8: Plot of a castanets "attack"

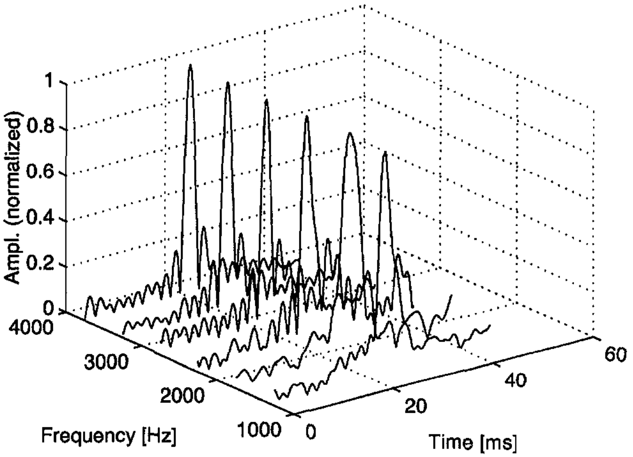


Figure 9: Normalized envelopes for castanets attack in different frequency bands

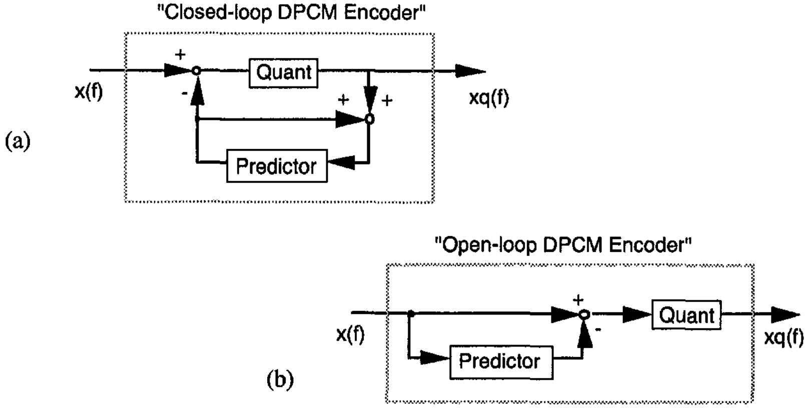


Figure 10: (a) Closed-loop and (b) open-loop prediction DPCM encoding schemes

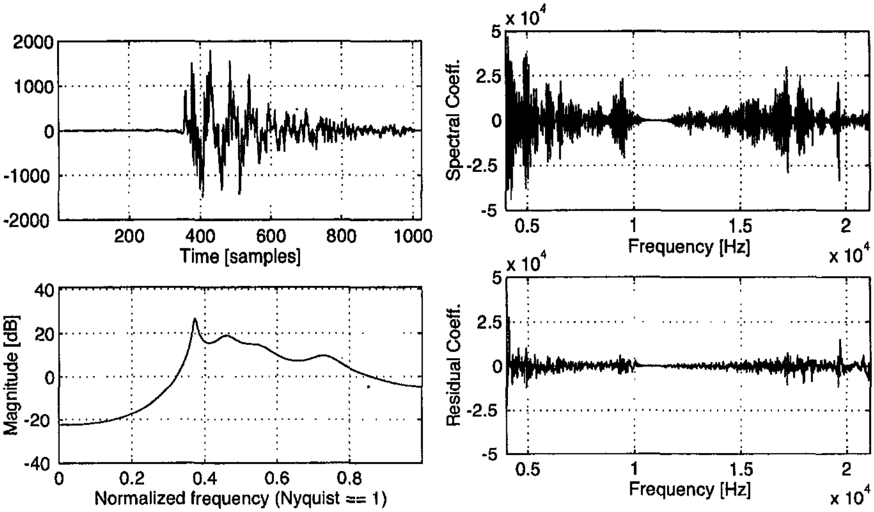


Figure 11: Transient input signal and resulting frequency response of TNS synthesis filter (left top and bottom), spectral coefficients and residual spectrum (right top and bottom)

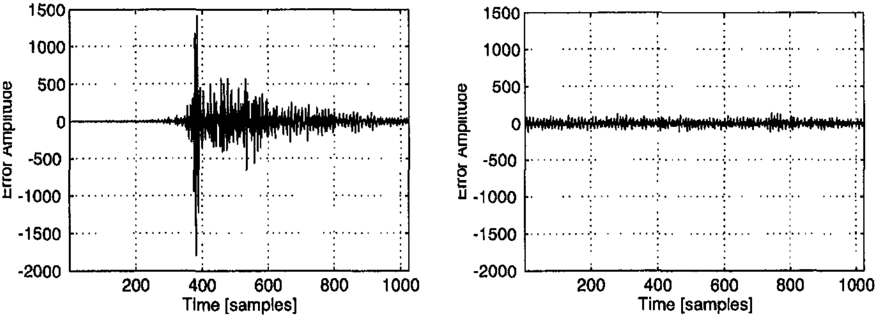


Figure 12: Resulting coding noise in decoded signal with (left) and without (right) Temporal Noise Shaping

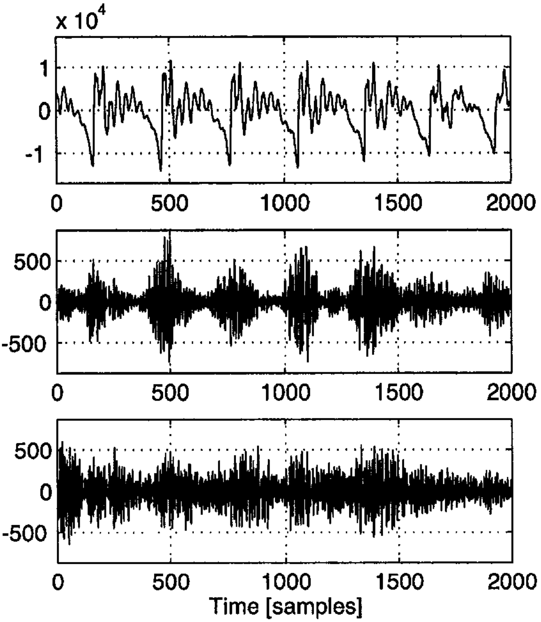


Figure 13: Original signal and coding noise with and without TNS (from top to bottom, excerpt from “German Male Speech”)

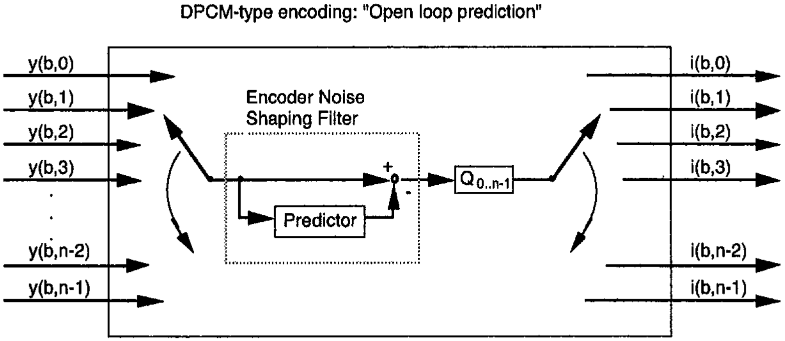


Figure 14: Replacing the PCM quantization / coding circuit by a DPC scheme

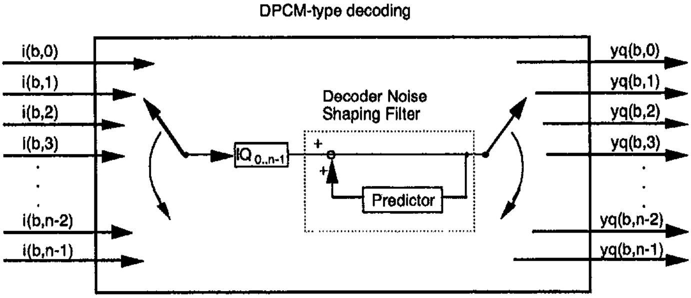


Figure 15: Replacing the inverse PCM quantization circuit by a DPC scheme

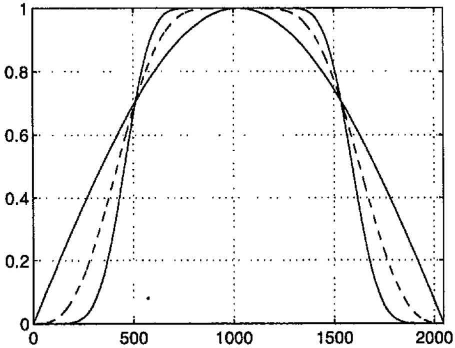


Figure 16: Three TDAC windows with different degree of overlap

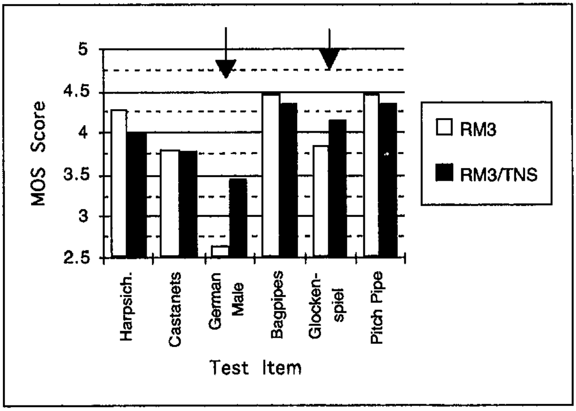


Figure 17: Test results from the ISO/MPEG2-Audio core experiment