Audio Engineering Society

Convention Paper

Presented at the 111th Convention
2001 September 21–24 New York, NY, USA

Perceptual Audio Coders "What to listen for"

Markus Erne, Vice President
Scopein Research
CH-5000, Aarau
Switzerland
Members of the AES-Technical Committee on Audio Coding

Markus.erne@scopein.ch
TC_CAS@aes.org

ABSTRACT

Low-bit rate audio coding has become a widely used technology during past years. By of the use of sophisticated signal processing techniques, exploiting psychoacoustic phenomena, nontransparent coding results in artifacts sounding very different from traditional distortions which are frequently not obvious at all to the untrained listener. The AES Technical Committee on Audio Coding therefore has started an activity to produce a CD-ROM which presents some of the most common coding artifacts in more detail. The CD-ROM not only explains and comments each of the coding artifacts separately but for each artifact, audio examples are presented, using different degrees of distortion, varying from "subtle" up to "obvious".

INTRODUCTION

Audio experts got trained in listening to distortion or signal degradation in analog systems over a period of more than 40 years. Digital technology however and low bit-rate coding especially has evolved during the last 10 years, thanks to the standardization activities within the ISO-MPEG group.
Audio Coding artifacts differ in their nature and their audibility much from well known audio distortions (wow&flutter, tape saturation, crosstalk, non-linearity, group delay distortion, intermodulation etc.) and they may be very difficult to identify in a networked audio system, being a cascade of different signal processing devices. The AES Technical Committee on Audio Coding started a project in year 2000 in order to produce a CD-ROM which demonstrates some of the most commonly known artifacts of perceptual audio coding algorithms.
The AES-TC on Audio Coding not only had to find appropriate, copyright-released audio material but it was the ultimate goal to keep the demonstrated artifacts as independent as possible from any commercially available coding system.
The coding artifacts which are explained and demonstrated using audio examples include: Pre-Echo artifacts, Speech-Coding artifacts, Binaural Masking Level Difference artifacts, Loss of stereo image artifacts, High-frequency limitation artifacts, Aliasing artifacts, artifacts due to

tandem coding and artifacts resulting from coding music with a speech codec.
During the release-phase of the project, the author noticed that there is not only a huge interest in such projects but additionally, most of the coding artifacts could never be explained as precisely without having a possibility to listen to audio files, explaining the phenomena much better than hundred pages of written text. This project additionally may also be a motivation to other technical committees and institutions to take up the idea of producing an educational CD-ROM and to start similar projects.

1. PRINCIPLES OF PERCEPTUAL AUDIO CODING

Before starting to discuss the different artifacts, a short review of perceptual audio coding will be provided. Based on the notion of the *critical bands* [4] [5] [6], the cochlea, from a signal processing perspective can be viewed as a bank of overlapping bandpass filters. It should be mentioned that the magnitude responses are asymmetric and non-linear because they are both level-dependent.

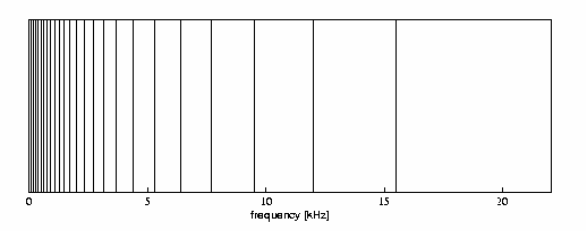

Fig 1.1. Critical bands

As it can be easily noticed, the band-splitting is non-uniform, showing constant bandwidth filters up to around 500 Hz and constant-Q filters above 500 Hz.
Simultaneous masking is a frequency domain phenomenon where a low-level signal (the maskee) can become inaudible (masked) by a simultaneously occurring stronger signal the (the masker), if masker and maskee are close enough in frequency [7].
Such masking is greatest in the critical band where the masker is located, and is effective to a lesser degree in neighboring bands. A masking threshold can be measured below which the low-level signal will not be audible. This masked signal can consist of low-level signal contributions, quantization noise, aliasing distortion, or transmission errors. The masking threshold, in the context of source coding also known as threshold of just noticeable distortion (JND) [8], varies in time. It depends on the sound pressure level (SPL), the frequency of the masker, and on the characteristics of masker and maskee. The slope of the masking threshold is steeper towards lower frequencies, i.e., higher frequencies are more easily masked. Additionally, the distance between masker and masking threshold is smaller in noise-masking-tone experiments than in tone-masking-noise experiments, i.e., noise is a better masker than a tone.

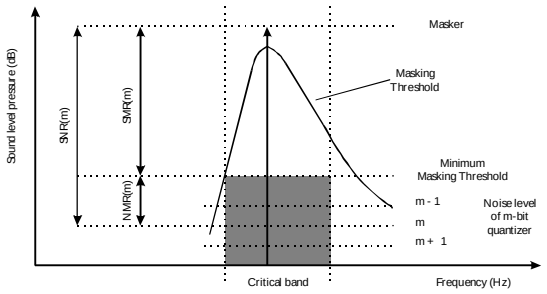

Fig 1.2. Influence of simultaneous masking on the signal-to-mask and noise-to-mask-ratio

Defining SNR(m) as the signal-to noise ratio resulting from an m-bit quantizer, the perceivable distortion in a given subband is measured by the noise-to-mask-ratio, NMR:

NMR(m) = SMR – SNR(m) in dB

Music and speech signals consist of many maskers, being tonal or noise-like, each of it having its own masking threshold, and all of them together will add up to the global masking threshold. In addition to frequency domain masking, the time domain phenomenon of temporal masking plays an important role in auditory perception. It may occur when two sounds appear within a small interval of time. Depending on the individual sound pressure levels, the stronger sound may mask the weaker one, even if the masked precedes the masker. (Premasking)

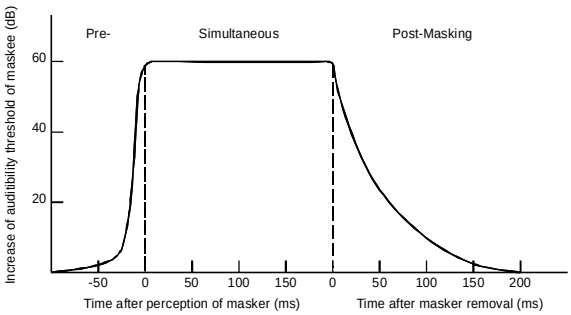

Fig 1.3 Temporal masking

Pre- and Postmasking can be exploited in perceptual coding algorithms and they become important when we start talking about block-artifacts (Pre-Echoes) in coding systems.

Hence the basic target of any perceptual coding scheme is to shape the quantization noise dynamically and adaptive to the signal in such a way that it always should be below the masking threshold. Different methods for achieving this goal have been suggested but as we will present in Section 9, coders, optimized for speech signals are not very suitable for audio signals. This automatically introduces us to subband coding and transform coding systems, although the differentiation between the two categories is mainly due to historical reasons. The idea is to split the source spectrum

into frequency bands in order to generate nearly uncorrelated spectral components in the subbands which then can be quantized and entropy-coded for each subband individually. In the encoder, an analysis-filterbank consisting of M-bandpass-filters is used in order to split the source spectrum accordingly.

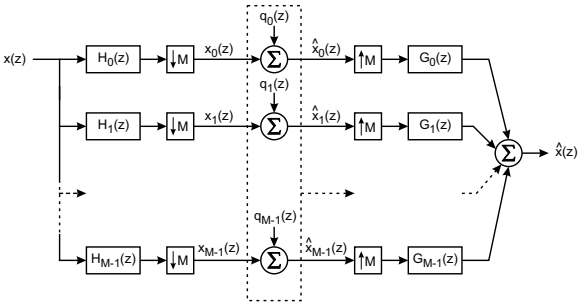

Fig 1.4. Subband Coding System with Quantizers

In the decoder, the set of subbbands is recombined, using a synthesis-filterbank. Normally, each filter is critically sampled, i.e. sampled at twice the bandwidth of the bandpass filter. Unfortunately ideal “brickwall” bandpassfilters do not exist and therefore aliasing cancellation techniques are being used in such filterbanks. Aliasing components between adjacent bands will cancel, due to the orthogonality between the analysis- and the synthesis filter. In Polyphase [9] and PQMF-filters [10], frequency domain aliasing cancellation is applied whereas in Transform-coders (MDCT, MLT) [11], time-domain aliasing cancellation is used.

Although, aliasing cancellation may allow perfect reconstruction under some conditions, each non-linearity i.e. the subband quantizer will degrade the alias cancellation process.

Aliasing artifacts will be discussed later in this paper, in Section 3.

In transform coding, a block of samples is linearly transformed via a discrete transform into a set of nearly uncorrelated transform coefficients. Block sizes ideally are chosen as large as possible in order to code almost stationary signal most efficiently and in order to get a reasonable ratio between the amount of side information (scaling factors, bit allocation data, ancillary data) and the amount of subband data. Nevertheless, a phenomenon, known as pre-echo can cause disturbing artifacts and will be discussed in Section 2.

Besides the block-size, the number of filter-bands is an important parameter in perceptual coding algorithms. Almost stationary signals require the use of large transform lengths, or filter banks with many subbands. But again, due to the Heisenberg principle [12], a large frequency resolution implies a poor temporal resolution and vice versa. Especially for speech signals, where voiced, unvoiced segments may alternate with fricatives, the use of long transforms or long filterbanks, having large frequency selectivity may introduce some “echoiness” into the coded speech signal. This artifact, known as “speech-reverberation” will be discussed in Section 5.

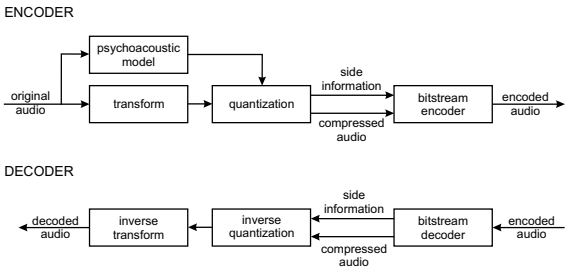

Fig 1.5. Encoder and Decoder of a perceptual subband-coding system

A lot of the secrets behind the implementation of a perceptual coder are related to the complexity of the perceptual model, controlling all the different coding blocks. The bit-allocation, for low bitrates and complex signal may be momentarily forced by the perceptual model to quantize some subbands (especially at high frequencies) to zero. A high frequency component of an instrument, contributing to the timbre of the instrument may therefore momentarily disappear and become audible again, when the bit-allocation of a following block has changed. This artifact which is called “birdies” will be explained in Section 4.

Another interesting phenomenon can be observed at low frequencies where the masked threshold can sometimes be lower, when listening with two ears rather than with one. This artifact, known as Binaural Masking Level Difference (BMLD) will be presented in Section 6. It can be summarized as follows: the detection of a signal in noise is improved when either the phase or the level differences of the signal at the two ears are not the same as of the masker.

Spatial perception in general is something which is far from being understood and some of the spatial cues which are provided in the literature [13] will provide a first insight to spatial hearing and localization of sound. The use of intensity stereo coding techniques provides a high potential for the saving of bits but may create some stereo imaging artifacts which are discussed in detail in Section 7.

Last but not least, there are a lot of system issues involved when using perceptual coders. The overall delay (determined by the length of the signal-transform, the maximum blocksize, the use of look-ahead-technology for block-switching as well as features like “bit-reservoir” jointly with the DSP-architecture and parallel programming), all will determine the overall system delay of an encoder-decoder configuration.

Problems certainly become worse when perceptual coders are connected in tandem. Since most processing (equalizing, editing, mixing) still is carried out in linear PCM-technology, each processing stage will require an additional decoding/re-encoding cycle. It is obvious, that quality will degrade with each cycle unless some new ideas like transcoding, presented in Section 8 will be used. For more details on the principles of perceptual coding, the reader can refer to [1], [2], [3].

2. PRE-ECHO ARTIFACTS

Typical blocksizes used in audio coders may range from 400..2048 samples per block. For each block, the subband-coding filterbank which can be realized using a transform (MDCT, MLT, ELT, WT) is computed and the masking threshold is normally computed using an additional FFT for increased frequency resolution.The reason for processing long blocks of audio is based on the fact that necessary side-information creates a smaller overhead for long blocksizes and additionally, slow-varying (or almost stationary) signals can be coded more efficiently. Nevertheless the block-size will directly influence the overall coding delay and therefore may become an important parameter for applications where a duplex communication is desired.

What is the problem of block-processing:

Suppose that during the current audio block which is coded, a time-domain transient occurs (e.g. a castanet signal). Because there is substantial signal energy in the attack, the perceptual model will allocate only a few bits to the quantizers in the subbands because a transient signal in the time-domain will spread out in frequency over many subbands and additionally, quantization noise could be masked during and after the transient due to spectral and temporal (post-) masking. In the decoder, the subbands samples are re-quantized and the permissible quantization noise which was supposed to be fully masked, now spreads out in time over the complete block and therefore will also precede the time-domain transient. This quantization noise which precedes the transient will cause audible time domain artifacts because it almost "announces" the transient in advance. While subband signals may be coarsely quantized using only a few bits, the quantization error in the subbands can not be considered as being uncorrelated with the signal itself.

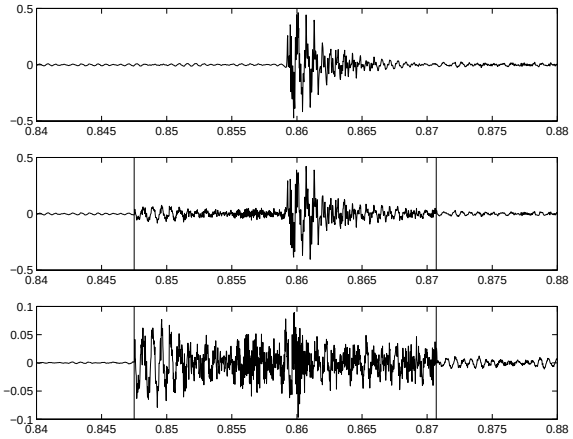

Fig. 2.1. castanet signal-original (top), the same signal after decoding (middle) and the difference between the two signals (bottom)

The transient of the castanet occurs in the middle of the current block and therefore a quantization error which may correlated with the transient signal itself due to coarse quantization can be noticed prior to the signal attack as so-called "Pre-echo, an artifact, very similar to the “copying effect” on analog tape.

It can easily be noticed that the quantization error preceding the attack may cause audible artifacts whereas the quantization error during the attack will be masked by the energy of the signal attack.

How can pre-echo be avoided ?

As discussed earlier, short block sizes create a large overhead for the side-information and do not allow to take profit from long-term stationarity in slow varying signals. But if the block is very short, the effect of the pre-echo distortion becomes smaller and the pre-echo artifacts may be masked due to temporal pre-masking. Therefore more advanced audio coders make use of a technique which is called adaptive block-switching [14]. A look-ahead of the energy build-up in the next block will allow to detect transients and will allow to switch to a shorter blocksize. For almost stationary signals, long block-sizes (as long as the maximum delay permits) are used. If a transient is detected, the blocksize will be switched to a shorter blocksize in order to avoid pre-echoes. For audio coders using lapped transforms, the window-size and optionally the shape of the time-window can adapt jointly with the smaller block-size. The price to pay is a less efficient coding scheme, because if a lot of transients occur, a large percentage of the available bit-rate will be used for side information during the small block mode.

3. ALIASING ARTIFACTS

Aliasing is well known from the PCM-sampling theorem, which states, that the sampling frequency must be at least twice the bandwidth of the signal to be sampled.

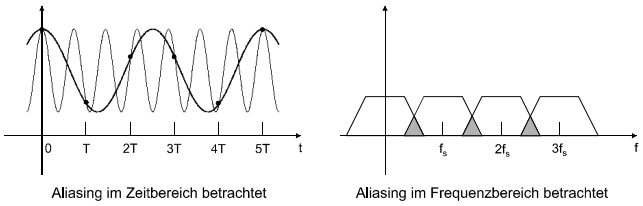

Fig.3.1. Aliasing in the time- and frequency domain

As already discussed, subband coders require a set of bandpassfilters and although these filters may be linear phase FIR-filters, their passband-ripple and their stopband attenuation will be limited. PQMF-filters have been used in

the MPEG1-standard [15] and they can be derived from a single prototype filter which then in its modulated version forms a filterbank of equally spaced bandpassfilters. Although these filters are based on aliasing cancellation [10] perfect reconstruction will not apply, when a quantizer, having only a few quantization steps is connected between the analysis and the synthesis part of the filter-framework. In practice however it turns out that due to the length of the filters of 512 taps, the artifact will hardly be audible under normal conditions.

Another approach has been choosen for the **Time Domain Aliasing Cancellation** filterbank, known as TDAC-Filterbank.

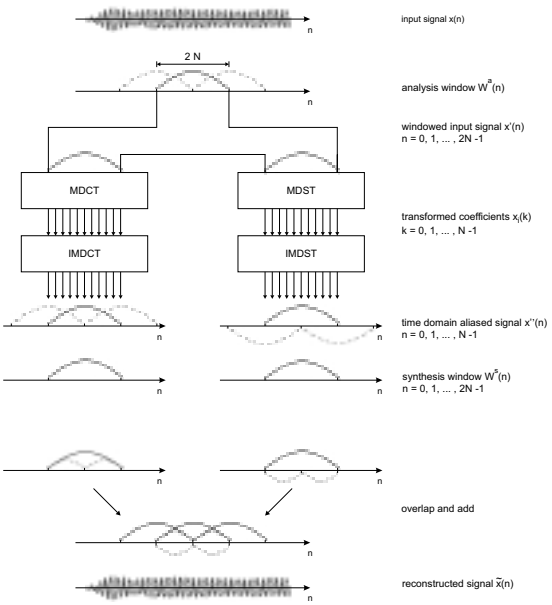

Figure 3.2. Principle of the TDAC filterbank

The time-domain signal is windowed, using a window of twice the length N , the number of subbands, and with a 50% overlap between successive blocks. The windowed signal is transformed using a discrete Sine Transform (DST) and a discrete Cosine Transform (DCT) where, after the inverse transforms correspondingly, the reconstructed time-domain signal will contain aliasing distortion. The aliased terms which for, easier explanation, are shown (dashed line) separately from the signal are time-reversed. Using a synthesis window and an overlap-add approach, these aliased terms will cancel and perfect reconstruction can be achieved. But again, subband-processing and coarse quantization may influence the cancellation process and audible time-aliases may appear, especially prior to the signal.

Depending on a frequency offset in the transform kernel, a oddly stacked time domain cancellation filterbank, OTDAC or MDCT or an evenly stacked time domain aliasing cancellation filterbank ETDAC or MDST will result. Malvar [11] unified the concept of time domain aliasing cancellation under the framework of lapped transforms (LOT) and showed that the oddly stacked MDCT can be considered as a Modulated Lapped Transform or MLT.

The MLT is widely used in audio coding, mainly in MPEG1&2 for Layer3, in Dolby AC-3, in Advanced Audio Coding (AAC) and in the Windows Media Player.

Wavelet Filterbanks:

Wavelets have become very popular since the 80's because they offer a mathematical framework for the design of non-uniform and signal-adaptive time frequency decompositions. Wavelets are based on a multi-scale approach, where at each scale in a tree structure, the frequency resolution is doubled and the time-resolution is halved. This very nice feature highly satisfies the requirements of music signals. Notes, played at low frequencies require a high frequency resolution in order to be separated, but as the mass of a low frequency string is bigger than for high frequencies, low frequency notes are never played in extremely fast tempos. High frequency notes on the other hand can be played very fast ("triller") and therefore require a high temporal resolution in order to be represented accurately. But, high frequency notes require a lower frequency resolution because the spacing of the notes is based on frequency ratios and hence a semi-tone in the fourth octave covers a much larger frequency range than at lower musical scales.

Nevertheless because wavelet-filterbanks are iterated filterbanks, where the individual filters can be considered as being in series, with each level of the wavelet-tree, the stopband-attenuation becomes worse.

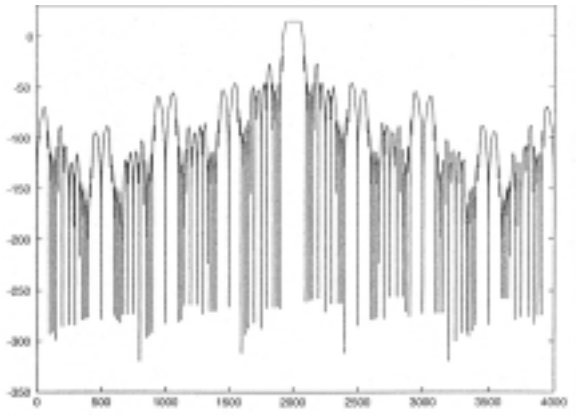

Figure 3.3. Selectivity of a Beylkin Wavelet-Filter of length 18

This is due to the fact that the temporal support of these wavelet-coefficients is only of the length of 10..50 and therefore basis functions of length 50 are being used for the approximation of the audio signal [16]. Using FIR-filters of length 50, only filters with limited stopband attenuation can be designed, and aliasing may appear in the sidelobes of the filter in the stopband-regions. These sidelobes can be sufficiently apart from the passband in order to create

aliasing in frequency regions with very little signal energy and where aliased quantization noise may not be masked.

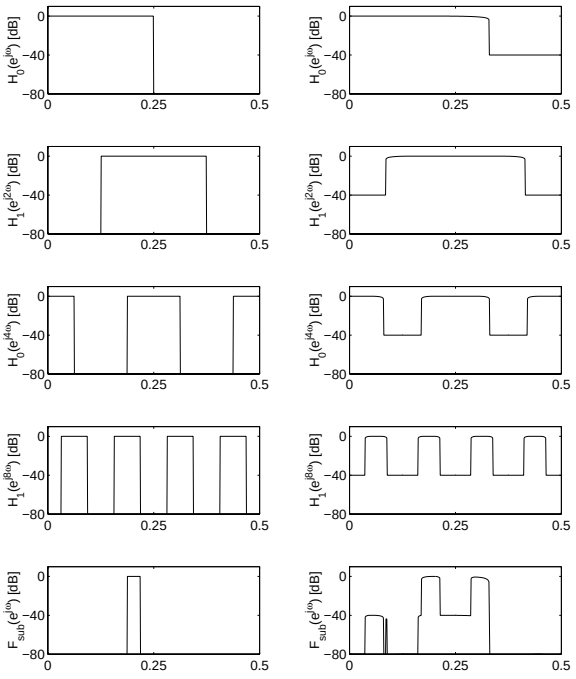

Figure 3.4. Comparison between the ideal (left) and the real (right) resulting frequency response of an iterated waveletpacket-filterbank

4. ARTIFACTS IN FREQUENCY AND THEIR VARIATION IN TIME (“BIRDIES”)

A perceptual audio codec includes, by definition, an algorithm, performing a model of the Human Auditory System[15]. Most perceptual coders compute a masking threshold estimation to determine the highest level of noise, which might be introduced at each frequency location but being imperceptible to the Human Auditory System. Other measures such as loudness or pitch modelling may also become part of the model but they are more related to the Auditory Interface (outer, middle and inner ear), without being directly connected to the higher level processing of the brain.

These high level processes seem to divide the sound event into a collection of independent items (*auditory objects*), and organizes them into several sets (*streams*) [13]; for example, a tune composed of both a violin and a flute playing at the same time would contain audio objects (every single note played) and a couple of streams, one for the notes played by the violin and other for the ones coming from the flute.

A lot these streams have been found during research in “Auditory Scene Analysis” [13] but we are still far from knowing, how the ear can develop its fantastic capability of re-grouping instruments from a mix of even hundreds of instruments.

Let us focus on **timbre**, being the whole set of characteristics that remains to be interpreted, once pitch, loudness and duration are extracted. The auditory streaming assembles some of the time-frequency components into a single auditory object, taking care of some complicated time-frequency relations between the time-frequency atoms. It is remarkable that isolated components can create new timbre and, hence new perceived objects can be created “simply” by changing timbre.

We already discussed the bit allocation algorithm which selects the best quantizer to be currently used, under the perceptual constraint, that quantization noise being inaudible. This bit-allocation procedure varies from block to block and therefore quantization noise is shaped on a block-by block basis. One of the possible quantizers is the “zero-quantizer” which will quantize the signal to zero.

Timbre vs. some typical bandwidth limitations

Timbre characteristics depend strongly on frequency, and on temporal structures, but their relationship may be extremely different. Some sound sources have their main characteristics in the first 4 or 5 KHz (i.e. speech), and most of their sounds are clearly recognizable with such a limited bandwidth (i.e. telephone). Of course high quality speech needs a higher bandwidth, up to 7 or 8 KHz, but there exist other types of sounds that need a much higher bandwidth and they simply would disappear if a limited bandwidth is being used. As the bit assignment varies from block to block, spectral coefficients may temporarily appear and disappear. The resulting change in timbre and these high frequency energy variations are known as “Birdies”-artifact and which have been reported in [17].

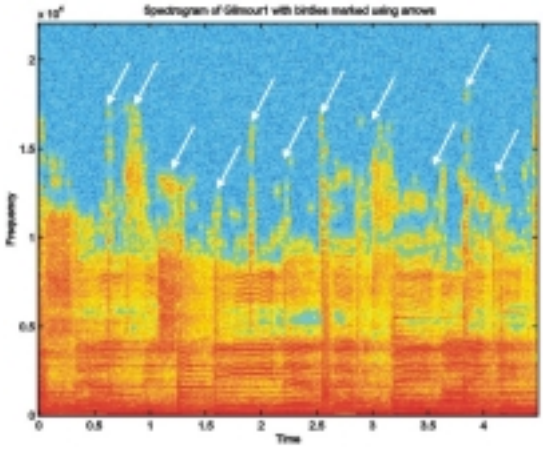

Figure 4.1. Spectrogram of a sound excerpt, indicating clearly the “birdies” artifact

One solution to avoid these types of artifacts might be to bandlimit the audio signal prior to coding. Although band-limitation will prevent the bit-allocation to spend more bits

at higher frequencies, there is still the possibility that these types of artifacts may occur.

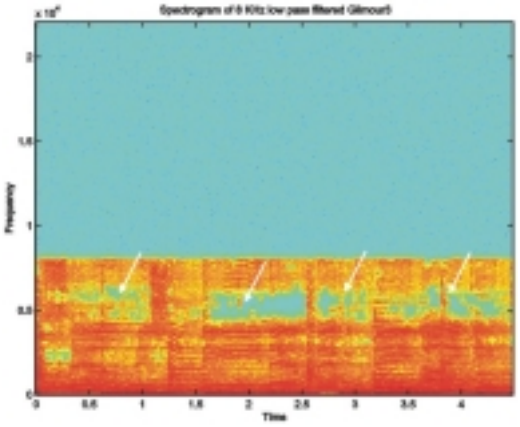

Figure 4.1. Spectrogram of a sound excerpt, indicating “birdies” artifact although the signal has been band-limited

5. SPEECH REVERBERATION ARTIFACTS

In order to achieve a high coding gain, large transform lengths or filterbanks with up to 2048 bands have been used in perceptual audio coders. For almost stationary and/or tonal signals, a large frequency resolution certainly will improve the efficiency of the coder because the input spectrum and therefore the masking threshold will have spectral peaks. A filterbank, having a large spectral resolution might isolate these spectral peaks, quantize them separately and therefore satisfy the condition that the quantization-noise in the neighbourhood of that spectral peak remains lower than the masked threshold.

Unfortunately, the Heisenberg uncertainty principle clearly indicates, that a large spectral resolution will imply a poor time resolution. Especially non-stationary signals (castanets) and speech-signals with some fricatives and plosives may momentarily require a higher temporal resolution. It turns out that these types of signals exhibit a broader spectrum, therefore generate a more flat masking curve and hence require a lower spectral resolution in order to be coded efficiently.

If the coder does not offer some adaptability in terms of time-frequency resolution, signal attacks and non-stationary signals will produce an artifact which is called “speech reverberation” artifact. This artifact becomes very obvious for speech signals using coding algorithms where large transformation lengths are used.

A possibility to overcome this problem is to switch the filterbank temporarily from a large spectral resolution i.e. 1024 bands to a lower frequency- but improved time-resolution.

In [18], a framework has been presented, offering an extremely flexible time-frequency tiling, completely adapting to the signal.

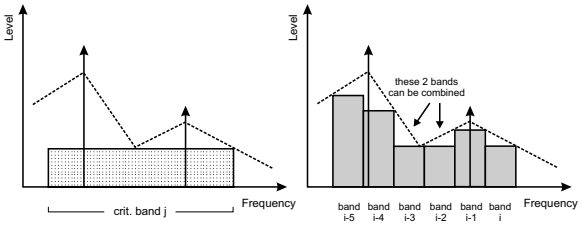

Figure 5.1. For a “peaky” input spectrum, a large frequency resolution has to be provided

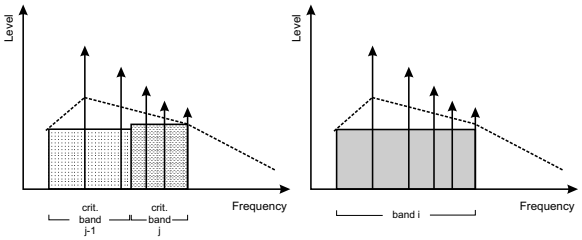

Figure 5.2. For a transient signal, having a flat input spectrum, a lower frequency resolution has to be provided

6. BINAURAL MASKING LEVEL DIFFERENCE

BLMD is a phenomenon that is observed at low frequencies when a masker and probe (masker is the signal doing the masking, probe the signal that is being masked) have specific time relationships. The classic experiment is as follows:
A narrowband noise masker of one critical bandwidth is presented identically at both ears (i.e. the same time signal), while the phase of the masked sinusoidal probe is alternated, i.e. presented in phase, and then out of phase, in the two ears. For the full effect to be noticed, this must be done at or below 500 Hz center frequency, although some effect has been noted to between 2 and 3kHz. In the case where both masker and probe are the same in both ears, a masking threshold very much like the single-ear masking threshold is observed, i.e. for the example given, the threshold of masking for the tone probe is approximately 5.5 dB.
In the case where the probe is out of phase, a difference between the signal with and without the probe is easily audible at this level. The experiment can also be run with the masker being applied in and out of phase, in which case release of the masking threshold is also observed, in some specific cases of up to approximately 20 dB.

In order to hear imaging artifacts in general, the listener must stop, at least temporarily, focusing on the usual range of artifacts, and try to allow the stereo signal to construct a soundstage, and then listen to artifacts and to position of things in the soundstage, noting both

omissions and commissions, i.e. new additions to the soundstage that are not present in the reference. An example is that of a commercial CD, that creates a situation with some coders whereby either channel sounds very similar to the original, but the STEREO signal sounds low-passey. The effect is not interference, cancellation, or something of that sort, but rather the flattening of the "pitch-like" signal envelop at high frequencies, causing the binaural hearing system to disregard the high frequencies ONLY IN THE STEREO PRESENTATION because the envelopes of the two channels are correlated neither to each other NOR to the envelope of the low frequencies in the signal.

7. LOSS OF STEREO IMAGE

For coding of high quality stereophonic (or multi-channel) audio signals at low bit rates, joint coding techniques have proven to be extremely valuable. On one hand they provide mechanisms to account for binaural psychoacoustic effects, on the other hand the required bit rate for the stereophonic signals may be reduced significantly below the bit rate for separate coding of the input channels. Currently, the most common joint stereo coding techniques are Mid/Side (M/S) stereo coding [19] and Intensity Stereo coding [20] [21].

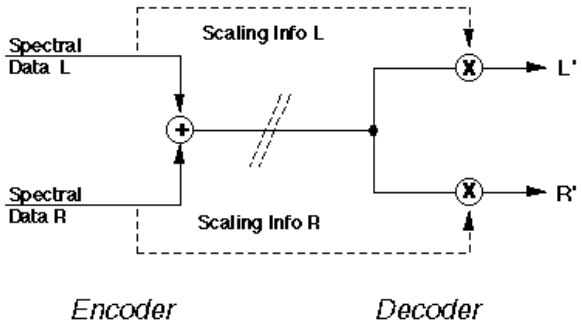

Figure 7.1. Principle of intensity stereo coding

Intensity stereo exploits the fact that the perception of high frequency sound components (e.g. above 4 kHz) mainly relies on the analysis of their energy-time envelopes [22] rather than the waveform itself. Thus, it is assumed sufficient to code the envelope of such a signal instead of its waveform. This is done by transmitting one common set of spectral coefficients ("carrier signal") that is shared among several audio channels instead of separate sets for each particular one. In the decoder, the carrier signal is scaled independently for each signal channel to match its original average envelope (or signal energy) for the respective coder frame. The scaling information is calculated and transmitted once for each group of spectral coefficients (scalefactor band). Effectively, the stereo image is recreated at the decoder side by a pan-pot-like operation for each spectral coder band.

As a consequence of the intensity stereo coding / decoding process, all output signals reconstructed from a single carrier are scaled versions of each other, i.e. they have the same envelope fine structure for the duration of the coded block (e.g. 10-20 ms). This does not present a major problem for stationary signals or signals having similar envelope fine structures in the intensity stereo coded channels.
For transient signals with dissimilar envelopes in different channels, however, the original distribution of the envelope onsets between the coded channels cannot be recovered.

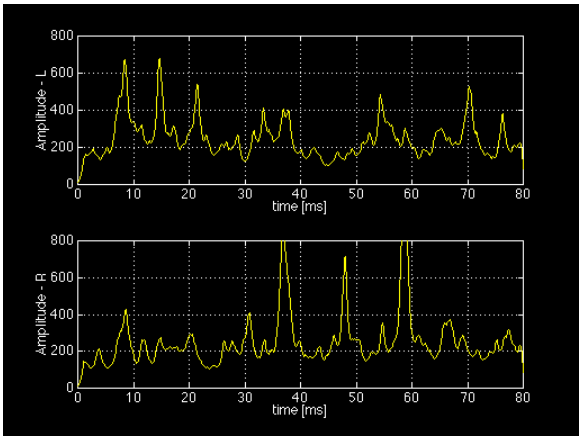

Figure 7.2. Excerpt from an “applause” item, showing left and right channel high frequency signal envelopes

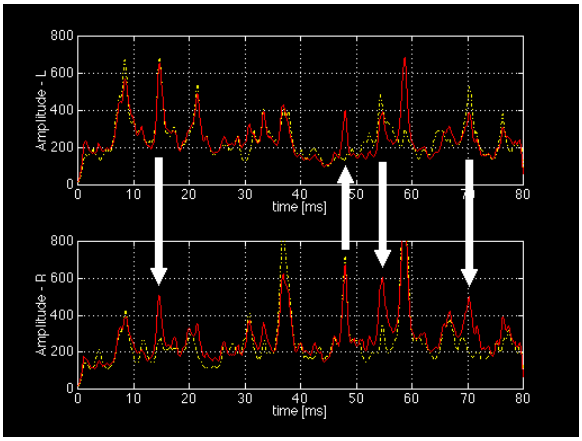

Figure 7.3.Excerpt from “applause” item, showing left and right channel high frequency envelope after intensity stereo encoding-decoding

8. TANDEM CODING

In each standard perceptual coder, the spectral representation of the input signal is altered slightly by the quantization process. The lower the bitrate, the coarser quantization is required to represent the signal within this given target bitrate. In this way, distortion is introduced which can be thought of as addition of quantization noise ("coding noise") which is shaped according to perceptual criteria, as estimated by the perceptual model.

With increasing deployment of low bitrate audio coding, use of audio compression can happen at various stages (contribution to production/studio, distribution between production facilities, emission/transmission of content to consumers, ...). Since processing/transmission of audio is still mostly done in uncompressed or even analog representation, this leads to repeated cycles of decoding, processing and re-encoding of the audio content. Similarly, a change of audio coding formats and bitrates usually requires a decoding/re-encoding cycle for format conversion.

Possible Solutions:

The most effective way of avoiding the accumulation of errors is to stay within the coded data format as long as possible. Thus, no further quantization processes introducing additional quantization noise are carried out. In fact, there is often no reason for leaving the coded domain and going via PCM, e.g. for copying purposes. Even when different algorithms are involved, "transcoding" (i.e. the conversion in the coded domain) can probably improve the outcome of tandem coding. If, however, further processing of the signal is required such as level change, equalization or reverberation, a return to the PCM domain is forced. Unfortunately, provisions for interfacing in the coded domain are not yet widely available today.

If decoding/re-encoding of the compressed audio content is necessary, it must be clear that a degradation in signal quality will happen. Thus, in order for the final coded audio to meet a desired target quality, the coding quality at intermediate coding steps must be significantly better than the target quality. In this way, quality losses due to tandem coding can be compensated by increased coding quality (and required bitrate) in intermediate coding steps.

Additional ideas might include: Inverse Decoder [23], use of MOLE-data [24] and ancillary data in digital interfaces.

9. AUDIO OVER SPEECH CODECS

Although historically, subband-coding techniques have been applied to speech coding, most speech coders are based on a model of the speech generation process of the human vocal tract. Linear Predictive Coding (LPC) [25] makes use of a time-varying all-pole filter whose excitations can be either an impulse train signal (tonal) or a noise generator (noise-like).

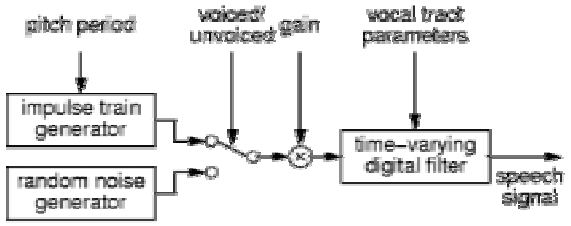

Figure 8.1.Speech Synthesis Model based on Linear Predictive Coding

It is clearly obvious that this system cannot synthesize tonal and noise-like components at the same time and therefore these type of coding structures will sound extremely un-natural when applied to broadband audio signals.

Code Excited Linear Prediction (CELP) [26] is another commonly used speech coding system based on an excitation codebook, a long-term prediction and a time-varying post-filter.

In general, speech coders offer a rather limited quality for wide-band audio signals, due to the following reasons:

- their underlying model is based on the generation of a single pitch period at a time and therefore cannot represent polyphonic music accurately
- all coding parameters are optimized for speech signals
- the shaping of the quantization noise is based on a very simple perceptual criteria, the (weighted mean square error)
- the coder has almost no adaptability in terms of time-frequency tiling, block-size etc.

In general it can be concluded that except for new concepts, based on parametric coding such as used in HILN, MPEG-4 Version 2 [27], coding schemes with simple models and using synthesis by analysis-techniques perform rather poor, compared to perceptual subband coding systems.

10. CONCLUSIONS

Although the interested reader might have got the impression that perceptual coding only will create artifacts, the author has to stress the point that during the past years, perceptual coding has gained a huge importance and made significant progress in terms of sonic quality. The deeper understanding on time-frequency relationships and the

progressing experience, using perceptual models has resulted in a family of coding schemes which for most applications are more than suitable and whose decoded signal can almost not be distinguished from the original, provided an appropriate bit-rate has been selected.

Nevertheless, a lot of questions had to remain unanswered throughout this paper. There is very little understanding on how humans have developed their absolute fantastic capability of discriminating and grouping sound events from a varying flow of air, produced by an orchestra or an instrument. Additionally, the principles of human hearing in general, the influence of the mental and physical condition on perception as well as the perception capabilities of ear-damaged human beings certainly are subject of further research.

It was the ultimate goal of this CD-ROM-project to provide universities, students, researchers, recording engineers, performing artists, musicians, publishers, journalists but also to the interested reader, with a tool, enabling him/her, to enter the real of audio coding. The project certainly is unique from the perspective that so far, never in coding history (and may be in the audio history in general), such a huge experience of the world’s best researchers in audio coding could be accumulated, resulting in a multimedia project, offering fundamental information, illustrations and many unique versions of audio demonstrations.

I would like therefore to express my deepest gratitude to all members of the Technical Committee on Audio Coding who contributed to this project, namely:

Jim Johnston
Jürgen Herre
Markus Erne
Karlheinz Brandenburg
Chris Lanciani
Gerald Schuller
Antonio Pena
Enrique Alexandre
Heiko Purnhagen
Marina Bosi

Many thanks to for reviewing the CD-ROM:

Martin Dietz, Bernd Edler, Grant Davidson, Louis Fielder, Mark Sandler, Mike Goodwin, Jens Spille, Peter G. Schreiner

And a special thanks to all the AES-officials, who supported this idea:

Roger Furness, Roy Pritts, Wieslaw Woszczyk, Robert Schulein, Patricia Macdonald, William McQuaide.

BIBLIOGRAPHY

[1] Brandenburg K., “MP3 and AAC explained”, *Proceedings of the AES 17-th International Conference on High Quality Audio Coding*, pp. 99-110, Florence, September 1999
[2] Bosi M., “Filter Banks in Perceptual Audio”, *Proceedings of the AES 17-th International Conference on High Quality Audio Coding*, pp. 125-136, Florence, September 1999
[3] Erne M., “Digital Audio Compression Algorithms”, *Proceedings of the First COST-G6 Workshop on Digital Audio Effects*, pp. 99-110, Barcelona, November 1998
[4] Greenwood D.D., “Critical Bandwidth and the Frequency Coordinates of the Basilar Membrane”, *J. Acous. Soc. Am.*, Oct. 1961, pp. 1344-1356
[5] Scharf B., “Critical Bands”, *Foundations of Modern Auditory Theory”*, Academic Press, New York, 1970
[6] Zwislocki J., “Analysis of Some Auditory Characteristics”, *Handbook of Mathematical Psychology*, John Wiley and Sons, New York, 1965
[7] Zwicker E., “Psychoacoustics Facts and Models”., *Springer Verlag*, 1990
[8] Jayant N.S., Johnston J.D., Safranek R., “Signal Compression based on Models of Human Perception”, *Proc. IEEE*, 81 (10), 1993, pp. 1385-1422
[9] Vetterli M., “Filter Banks allowing Perfect Reconstruction”, *Proc. IEEE SP*, Vol. 10, No. 3, April 1996, pp. 219-244
[10] Nussbaumer H., “Pseudo QMF Filterbanks”, *IBM Technical Disclosure Bulletin*, 24, 19981, pp.3081-3087
[11] Malvar H., “Signal Processing with Lapped Transforms”, *Artech House*, Norwood, 1992
[12] Blatter C., “Wavelets, eine Einführung”, *Vieweg*, 1998
[13] Bregmann A., „Auditory Scene Analysis“, *MIT-Press*, 1990
[14] Edler B., Coding of audio signals with overlapping transform and adaptive window functions”, *Frequenz, 1989*, 43, pp. 252-256
[15] ISO-IEC JTC1/SC29/WG11, “ISO/IEC 11172-3”, *ISO-IEC*, 1992
[16] Erne M., “Perceptual and Near-Lossless Audio Coding based on Signal-adaptive Wavelet Filterbank”, *AES 106-th Convention*, Preprint 4934, May 1999
[17] Pena A., “Técnicas de modelado psicoacústico aplicadas a la codificación de audio de muy alta calidad”, *PhD-thesis*, Universidad Politécnica de Madrid, 1994
[18] Erne M., “Signal Adaptive Audio Coding using Wavelets and Rate Optimization”, *PhD-thesis*, Swiss Federal Institute of Technology, ETH, 2000
[19] Johnston J., Ferreira A., “Sum-Difference Stereo Transform Coding”, *IEEE ICASSP*, 1992, pp. 569-571
[20] Veldhuis R., v.d. Waal R., “Subband Coding of Stereophonic Digital Audio Signals”, *IEEE ICASSP*, 1991, pp. 3601-3604
[21] Herre J., Brandenburg K., Lederer D., „Intensity Stereo Coding“, *AES 96-th Convention*, Preprint 3799, 1994
[22] Blauert J., “Spatial Hearing”, *MIT Press*, 1983
[23] Herre J., Schug M., “Analysis of Decompressed Audio – The Inverse Decoder”, *AES 109-th Convention*, Preprint 5256, 2000
[24] Fletcher J., “ISO/MPEG Layer2 – Optimum Re-Encoding of Decoded Audio using a MOLE Signal”, *AES 104-th Convention*, Preprint 4706, 1998
[25] Makhoul J., “Linear Prediction: A Tutorial Review”
[26] Atal B., Schroeder M., “Stochastic Coding of Speech Signals at Very Low Bit Rates”, *Proc. IEEE Int. Conf. on Communications”*, May 1984
[27] Purnhagen H., Edler B., “Error Protection and Concealment for HILN MPEG-4 Parametric Audio Coding”, *AES 110-th Convention*, Preprint 5300, 2001