A guideline to audio codec delay

Manfred Lutzky[1], Gerald Schuller[2], Marc Gayer[1], Ulrich Krämer[2] and Stefan Wabnik[2]

[1] Fraunhofer Institute for Integrated Circuits IIS, Am Wolfsmantel 33, 91058 Erlangen, Germany

[2] Fraunhofer Institute for Digital Media Technology IDMT, Langewiesener Strasse 22, 98693 Ilmenau, Germany

Correspondence should be addressed to Manfred Lutzky (ltz@iis.fraunhofer.de)

ABSTRACT

Digital audio processing has been revolutionized by perceptual audio coding in the past decade. The main parameter to benchmark different codecs is the audio quality at a certain bit-rate. For many applications, however, delay is another key parameter which varies between only a few and hundreds of milliseconds depending on the algorithmic properties of the codec. Latest research results in low delay audio coding can significantly improve the performance of applications such as communications, digital microphones, and wireless loudspeakers with lip synchronicity to a video signal.

This paper describes the delay sources and magnitude of the most common audio codecs and thus provides a guideline for the choice of the most suitable codec for a given application.

1. INTRODUCTION

Audio coders like MP3 or MPEG-AAC have become common in applications like storage or broadcast. On the other hand, speech coders are used for delay-critical communications applications at low bit-rates, but also low quality. Audio coders usually use subband coding, since this principle allows for a straightforward inclusion of psycho-acoustic models. The more subbands are used, the higher the compression ratio can be. The goal of high compression ratios lead to audio coders with a high number of subbands. For instance the MPEG-AAC coder has 1024 subbands, switchable to 128 bands. This structure leads to a high encoding/decoding delay, which makes systems like this unsuitable for communications applications. For that reason the MPEG-AAC Low Delay coder was developed. It obtains a lower delay by having a reduced number of subbands (480 instead of 1024). The down side is that this reduced number also leads to a somewhat reduced compression efficiency.

A different coding principle is used by speech coders,

namely predictive coding. Here the highest compression performance can be obtained by the longest predictors. But unlike subband coding, predictive coding does not lead to higher delays when higher compression performance is the goal. That property makes this principle also attractive for low delay audio coders. It enables audio coders with a very low delay and still a state-of-the-art compression ratio. An example is the low delay audio coder described e.g. in [1].

2. SOURCE OF DELAY

2.1. Filter bank

The filter banks used in audio and speech coding are usually in 2 classes. One is the class of MDCT filter banks, the other class is QMF filter banks. MDCT filter banks are used for instance in the MPEG 1/2 coder Layer 3, in the MPEG-AAC coder, the g.722.1 wideband speech coder, and many other similar coders. MDCT filter banks have a perfect reconstruction (no error if analysis and synthesis are directly in cascade), and their filter length is between 1 to 2 times the number of subbands, usually 2 times the number of subbands.

QMF filter banks are used in the MPEG 1/2 coder Layers 1 and 2, or the G.722 speech coder. QMF filter banks don't have perfect reconstruction, but near perfect reconstruction, their reconstruction error is very small. Their filter length is usually much longer than 2 times the number of subbands. For that reason they are usually used in applications with only a few subbands.

Both classes have in common, that they are so-called orthogonal filter banks (their polyphase matrices are unitary). This has the consequences for the system delay of the filter banks. The system delay is the delay of a signal through the analysis filter bank and the synthesis filter bank. This delay results from the downsampling, which is conducted in the analysis filter bank, and from the shape and length of the filters. Orthogonal filter banks automatically have a shape of the filters such that the system delay equals the length of its filters minus 1. Hence in this important case the system delay is very easy to determine.

$$N_{filte\_bank} = length\_of\_filters - 1$$

As an example: the MPEG-AAC audio coder has an MDCT filter bank with 1024 subbands. Hence its system delay is 1023 samples.

2.2. Block switching

Audio coders often use block switching to avoid or reduce so-called pre-echo artefacts. This block switching is a switching of the filter bank to a lower number of subbands and hence a higher time resolution [2]. In audio coding those block switching is only common for the MDCT filter banks. Using orthogonal MDCT filter banks, this switching cannot be done immediately, but it needs a so-called transition window in one block. In order to start this sequence for the switching in time before transient signals, like percussive sounds or "attacks", a look-ahead into the audio signal is necessary, which corresponds to a delay. In general, an encoder using block switching incurs an additional delay:

$$N_{look\text{-}ahead} = \frac{frame\_size\_long}{2} + \frac{frame\_size\_short}{2}$$

where $frame\_size\_long$ and $frame\_size\_short$ represents the number of subbands for long and short blocks (e.g. 1024 and 128 for MPEG-2 AAC) [3].

Note: The use of short blocks causes an overhead of side information bits. Also, the lower number of subbands often leads to a reduced compression performance and hence a corresponding peak in bit-rate. Hence a well thought out block switching strategy switches down rather seldom to limit the related bit-rate peaks. These short bit-rate peaks can be compensated by the use of bit-reservoir technique which introduces an additional delay. This is explained in a chapter.2.4.

2.3. Prediction

Predictive coding is used in 2 classes. One class is block wise prediction; the other class is backwards prediction. In block wise prediction a block of data is analysed and the most suitable prediction coefficients are computed in the encoder and send as side information to the decoder. Because the block of data is needed before encoding the data, the block length is the minimum delay of this scheme.

In backwards prediction, the predictor coefficients are updated at each sample, but only based on past samples. This has the advantage that no side information needs to be transmitted to the receiver, since the receiver also has the past samples. Since the predictor update is based only on past samples, it also does not introduce any delay. Also the prediction itself introduces no delay, the linear system with the predictor can be viewed as a minimum phase system.

Altogether, this system provides compression at zero delay, which makes it very attractive for delay-critical applications. Fig. 1 and Fig. 2 show the principle of predictive coding.

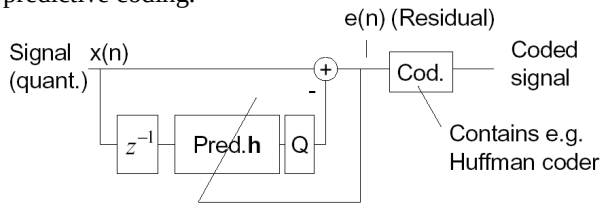

Fig. 1: Predictive encoder

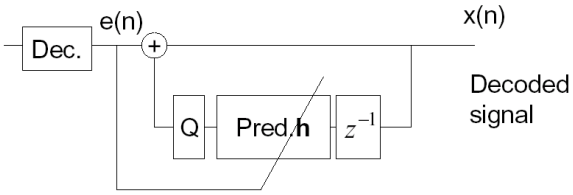

Fig. 2: Predictive decoder

2.4. Bit reservoir

Since not all segments of an audio signal are equally demanding to code, the bit reservoir technique has been introduced to several constant bit rate codecs in order to adapt the number of available bits to the signal characteristic. Since the use of the bit reservoir is equivalent to a local variation in bit rate, the size of the input buffer of the decoder must be adapted to the maximum local bit rate (i.e. the maximum number of bits which can be allocated for a single frame per channel). In fact, the overall delay of the audio coder may be dominated entirely by the size of the bit reservoir. The delay expressed in terms of samples caused by the bit reservoir is

$$N_{bitres} = \frac{bitres\_size}{bitrate} * F_S$$

where bitres-size is the bit reservoir size expressed in bits and F, is the sampling rate in Hz.

The bit reservoir technique can be realised with variable frame length or fixed frame length.

2.4.1. Fixed frame length:

Each transmitted block has the same number of bits. The coded audio data are inserted before or directly after this header. The position relative to the header can vary. In the case the bit reservoir is empty the coded data will start directly after the header. There is a so-called back pointer transmitted with the side information that contains the starting point of the encoded audio signal relative to the header. The bit reservoir delay occurs in the encoder. The decoder can immediately start the output of decoded audio samples after one complete frame is arrived.

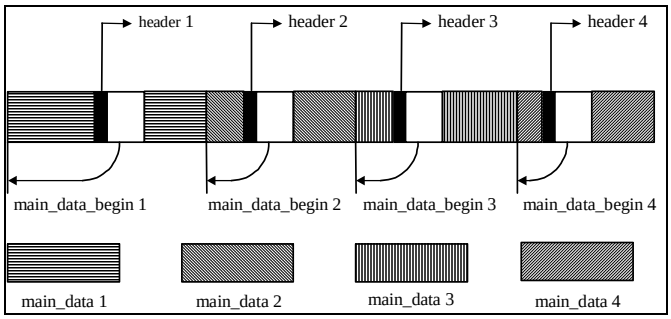

Fig. 3: Fixed frame length

2.4.2. Variable frame length

All coded audio data is directly transmitted after the header. Therefore an encoder can start to transmit one frame immediately after encoding it. The delay is shifted to the decoder that has to wait until one complete frame plus the current state of the bit reservoir has arrived.
Note: The bit reservoir delay of a variable frame length codec can be avoided in the case of packet oriented transmission!

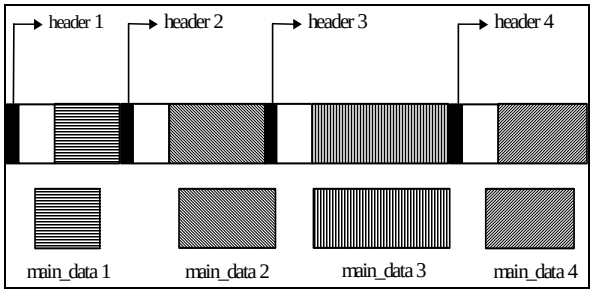

Fig. 4: Variable frame length

3. OVERVIEW AUDIOCODECS:

3.1. Subband coders

Perceptual audio codecs consist of the following basic building blocks (see Fig. 5) [4]:

- Filter bank, time frequency transform: In the encoder the audio input signal is decomposed into subsampled spectral components. Together with the corresponding filter bank in the decoder it forms an analysis/synthesis system.
- Perceptual model (encoder only): Using rules known from psychoacoustics, an estimate of the time and frequency dependent masking threshold is calculated from the output of the analysis filter bank and/or the time domain audio signal.
- Quantization and coding: The spectral components are quantized and coded with the aim of keeping the noise, which is introduced by quantizing, below the masking threshold. Depending on the sophistication of the algorithm this can include simple block companding up to analysis-by-synthesis systems with additional noiseless compression.
- Bitstream multiplex: The quantized and coded spectral data and side information, e.g. quantization step sizes, are assembled forming a bitstream.

In the decoder these building blocks are used in the opposite direction without the need for a perceptual model.

3.1.1. Philips Low-Complexity Subband Codec (SBC)

The Philips Low-Complexity Subband Codec was first presented at the 98th AES [6]. They developed a coding system with an algorithmic delay of 5 ms at a sampling frequency of 32 kHz. The targeted data rate for mono audio was 128 kpbs. These specs are the reason why the SBC was chosen as mandatory codec for the Bluetooth Advanced Audio Distribution Profile (A2DP) by the Bluetooth Special Interest Group (SIG).

Basically, the structure of the SBC is quite similar to most of the well known transform coders. A critically-sampled cosine modulated polyphase filterbank transforms the time domain signal into eight equally spaced subbands, which means 2 kHz per band. At 32 kHz, the delay of 2.5 ms corresponds to a filter length of L=80. The cascade of the analysis and synthesis filter-bank results in near perfect reconstruction.

The remaining 2.5 ms of delay are introduced by the block companding APCM quantization method. Eleven subband samples are grouped into a vector, resulting in a delay of ten subband samples. Each vector is normalized by a scalefactor. These scalefactors are used to determine the resolution of the quantizers in such way that the audibility of the introduced distortion is minimized. No entropy coding is applied, as the quantizers use just as many bits as necessary to fill 128 kpbs.

For the implementation in the Bluetooth A2DP the SBC was enhanced to allow multiple data rates. For this reason, the filter-bank can be reduced to four subbands and other sampling rates are possible. Furthermore, stereophonic signals can be coded with the A2DP version of the SBC, too [7].

3.1.2. MPEG audio coding standards

MPEG (ISO/IEC JTC1/SC29/WG11, better known as Moving Pictures Expert Group) started in 1988 to develop standards for the coded representation of moving pictures, audio, and their combination. This work has led to a set of ISO documents in which the following perceptual audio codecs have been standardized: MPEG-1/2 Layer-1/2/3, MPEG-2/4 AAC, MPEG-4 High Efficiency AAC, and MPEG-4 AAC Low Delay.

3.1.3. MPEG-1/2 Layer-2

In MPEG Layer-2 a polyphase filter bank (PQMF) creates 32 subband representations of the audio input signal [5]. The PQMF has filters of length 512 taps. The subband signals are then quantized and coded under the control of a psycho-acoustic model from which a block-wise adaptive bit allocation is derived. Layer-2 introduces further compression by redundancy and irrelevance reduction on the scale factors. A bit reservoir technique is not present. The basic frame length is 24 ms (at 48 kHz sampling frequency), corresponding to 1152 time samples. For Layer-2, encoder complexity is moderate and decoder complexity is low. A good quality bitstream requires around 96 kpbs/ch.

The resulting system delay of the filter bank is 511 samples. Since the coder transmits 36 filter bank blocks of 32 subbands in one frame, the delay is 32*(36-1)+511=1631 samples.

3.1.4. MPEG-1/2 Layer-3 (MP3)

Layer-3, better known by its nickname MP3, uses a hybrid filter bank consisting of the PQMF present in Layer-2 and a cascaded switchable modified discrete cosine transform (MDCT) to gain increased frequency resolution. The filter bank consists of an 18 subband MDCT for steady state signals and a 6 subband MDCT for transient signals. For 18 subbands the MDCT has a filter length of 36 taps, and for 6 subbands it has 12 taps. This results in a total of 32*18=576 bands in the steady state case, and 32*6=192 bands when it switches down. The basic frame length is 24 ms (at 48 kHz sampling frequency), corresponding to 1152 time samples [4][5], consisting of 2 filter bank blocks of 576 subbands.

For Layer-3, decoder complexity is low, while encoder complexity is high due to a sophisticated psycho-acoustic model, non-uniform quantization and adaptive segmentation and entropy coding of the quantized values as well as the block switching which chooses between the short and long MDCT mentioned above. This block switching introduces additional delay in the encoder since it requires a look-ahead to 384 future audio samples.

Furthermore Layer-3 uses a bit reservoir, and is a fixed frame length codec as described in chapter 2.4.1. This introduces additional delay independent from the transmission mode. For a detailed description of the Layer-3 bit reservoir size have a look at [19]. The bit reservoir delay in case of packet switched transmission can be reduced by the usage of a variable bit-rate mode or furthermore by embedding mp3 in MPEG4. Besides minimising mp3 delay combination all MPEG 4 features can be used such as combination with MPEG 4 video enhanced cutting mechanism and object oriented mechanism are possible [8] For Layer-3, a good quality bitstream requires about 64 kpbs/ch.

The resulting system delay of the MDCT filter bank is 35 samples plus a look-ahead delay of 12 samples, equals 47 samples. This is in the down-sampled domain after the PQMF filter bank. Hence in the audio domain it means 47*32 =1504 samples. The PQMF adds 511 samples delay. Since two filter bank blocks are transmitted together this adds another 576 samples, totalling 1504+511+576=2591 samples.

The size of the bit reservoir can be calculated with following formula:

$$\text{bitreservoir} = \max(0, \min(\text{limit}, 7680 - \text{frame\_len}))$$

where limit is 4088 for MPEG1 and 2040 for MPEG2 and MPEG2.5 and frame_len is the length of one bitstream frame

As the bit reservoir delay can only reach values of multiple of the length of main_data of a frame the max_bitreservoir value has to be rounded up to a integer of this value [19]. Some examples are mentioned in Table 1

3.1.5. MPEG-2/4 Advanced Audio Coding (AAC)

Among the various MPEG-2 AAC profiles [20] and MPEG-4 AAC audio object types [21] AAC Low Complexity (AAC LC) is used most often. AAC follows the same basic coding paradigm as Layer-3 but adds numerous improvements and coding tools for better audio quality at all bit-rates. The most important enhancements to mention here are improved joint stereo and Huffman coding, more flexible block switching to further reduce the amount of pre-echo artifacts, and the temporal noise shaping (TNS [12]) tool which controls the temporal shape of the quantization noise by applying a filtering process to parts of the spectral data.

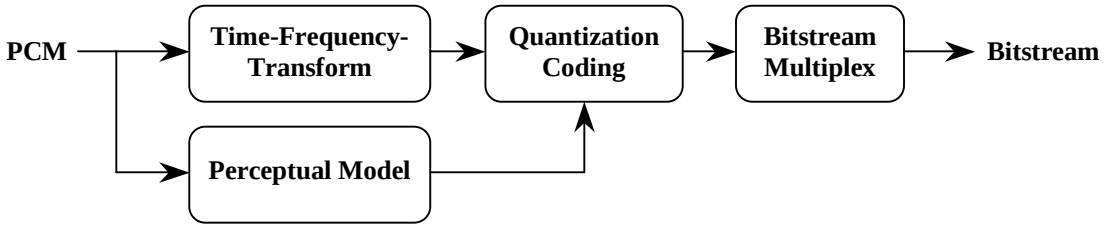

Figure 5: Basic structure of an MPEG perceptual audio encoder

AAC LC uses a 1024-subband MDCT with a filter length of 2048 taps for steady state signals, and switches to a 128-subband MDCT with a filter length of 256 taps for transient-like signals, resulting in a basic frame length of 1024 samples, or around 21.3 ms (at 48 kHz sampling frequency). Similar to Layer-3 the block switching look ahead and the bit reservoir technique add additional coding delay. The block switching look-ahead requires 576 audio samples. The system delay of the filter bank is 2047+576=2623 samples, or 54.6 ms at 48 kHz sampling.
The bit reservoir size per channel is 6144 bits minus the average frame length in bits, which is very significant in terms of delay. However, all AAC codecs are variable length codecs which means that the bit reservoir delay can be neglected if a packet-oriented transmission mode is chosen. For AAC LC, encoder complexity is high and decoder complexity is low with good quality bitstreams at around 48 kpbs/ch or below.

3.1.6. MPEG-4 High Efficiency AAC (HE AAC)

Recent work in MPEG-4 audio adds to the standard a backwards compatible combination of AAC and spectral band replication (SBR, [22]) which improves coding efficiency in particular for applications targeting low and very low bit-rates. In HE AAC the core codec (AAC) operates at half the audio sampling rate and the missing high frequency region is recovered in the decoder by the SBR module based on the transmitted lowpass signal and a small amount of control data which is extracted in the encoder before down sampling of the input signal.
The HE-AAC encoder consists of the AAC core encoder after a down-sampler by 2. In parallel there is a 64 subband QMF filter bank, analyzing the spectral content. In the decoder, there is the AAC core decoder operating at half the sampling rate, followed by a QMF analysis filter bank with 32 subbands and a filter length of 320 taps. For these 32 subbands, SBR generates the 32 high bands. For this generation it also needs a look-ahead of 6 subband blocks. The resulting 64 subbands are then fed into a 64 subband QMF synthesis filter bank to obtain the audio signal at the high sampling frequency. Hence the encoding/decoding delay at 48 kHz sampling rate is as follows. The AAC core coder has the 2623 samples delay, but now at 24 kHz sampling rate, or 109 ms. We neglect the down-sampling by 2 for the encoder. In the decoder we need to add the delay of the QMF filter bank, which is 288 samples (=319-31, as no blocking delay occurs) at 24 kHz sampling rate, or 12 ms. Since a look-ahead of 6 subband blocks is needed, this adds 6*32=192 samples or 8 ms delay. The total delay is 2623+288+192=3102 samples, or 129 ms (at 24 kHz sampling rate), which is also the delay for the entire coder if the down-sampling by 2 in the encoder is neglected.
For HE AAC, encoder complexity is high and decoder complexity is moderate. HE AAC provides good quality at bit-rates around 28 kpbs/ch. For high quality, the core coder AAC LC is used at higher bit-rates.

3.1.7. MPEG-4 AAC Low Delay

The MPEG-4 standard in its version 2 adds MPEG-4 AAC Low Delay (AAC LD), which targets applications where a low end-to-end delay is important, as for example in two way communication such as telephony or teleconferencing. In addition AAC LD is only standardized in an error robust form to cope with transmission errors often present in mobile communication scenarios. Various listening tests [3] show a comparable performance to MPEG-2 AAC, at a bit-rate about one third higher for AAC LD.
In AAC LD the original AAC codec has been modified in a number of ways to decrease the overall codec delay to 20 ms algorithmic delay and less than 40 ms in a hardware implementation. AAC LD uses a 480- or 512-subband MDCT, with filter lengths of 960 or 1024 taps, without block switching to avoid the delay introduced by the block switching look ahead. This also avoids high bit-rate peaks associated with the short block mode in other coders. Hence the system delay of the filter bank is 959 or 1023 samples. This corresponds to 20 ms or 21.3 ms at 48 kHz sampling rate.
The use of the bit reservoir is restricted to only a small number of bits (around 100 as used in Table 1) or completely avoided. The basic frame length of AAC LD is 480 samples or 10 ms at 48 kHz sampling frequency with a 480-subband MDCT. Both encoder and decoder complexity are comparable to AAC LC with an overhead of around 20% due to the higher number of audio frames processed per second and the rather inefficient filter bank implementation in case the 480-subband version. The audio quality of AAC LD is slightly better compared to mp3 at the same bitrate [3]

3.1.8. Delay considerations

Of the various possible delay sources described in

	Algorithmic delay without bit reservoir	Hardware, 100% enc workload, burst transmission	Hardware, 100% enc workload, continuous transmission	Hardware, 30% enc workload, burst transmission	Hardware, 30% enc workload, continuous transmission
MPEG-1 Layer-2 192 kpbs	34 ms	-	-	-	-
MPEG-1 Layer-3 128 kpbs	54 ms	118 ms	142 ms	107 ms	131 ms
MPEG-4 AAC 96 kpbs	55 ms	82 ms	211 ms	63 ms	192 ms
MPEG-4 HE AAC 56 kpbs	129 ms	184 ms	361 ms	145 ms	322 ms
MPEG-4 AAC LD 128 kpbs	20 ms	33 ms	44 ms	24 ms	35 ms

Table 1: Overall delays of various audio codecs operating at their typical bit-rates and at a sampling-rate of 48 kHz stereo.

previous chapters of this paper the class of perceptual audio codecs has to deal with delay introduced by the filter bank, block switching look-ahead, bit reservoir techniques, and post processing. These items contribute to the algorithmic delay of an audio codec. In a hardware implementation additional factors like processing time and transmission of the bitstream have to be considered. Of course, these heavily depend on the chosen processor architecture and transmission path.

Table 1 gives an overview on the various overall delays present in the perceptual audio codecs introduced in the previous chapters.

The “Algorithmic delay” is a best-case value including only delay sources introduced by the algorithm itself without bit reservoir. This would be equal to a hardware implementation where unlimited processing power and transmission speed is available.

In a hardware implementation the available processing power is limited and often the processor clock is reduced as far as possible to have lower power consumption. This leads to a scenario where the processor workload of the encoder is almost 100% causing an additional delay equal to the frame length of the codec. For the decoder the same processor clock as for the encoder with a workload of 30% is assumed, leading to an additional delay equal to 30% of the codec’s frame length. This overall delay type is called “Hardware 100% enc workload, burst transmission” in Table 1 and a burst-like, packet-oriented transmission of the bitstream is assumed where the bandwidth of the network is much higher than the bit-rate of the audio codec.

Many transmission paths however, do not offer this high bandwidth and their transmission clock is equal to or only slightly higher than the bit-rate of the audio codec. An example for that would be transmission via ISDN, an analogue telephone line or via DECT. In this case an additional delay equal to the average bitstream frame length divided by the network transmission clock is present. In Table 1 this is called “Hardware 100% enc workload, continuous transmission”.

In case the overall realtime delay should be further reduced it is advisable to increase the processor clock and thus decrease workload and the delay caused by processing the data. As an example one can assume a processor workload of 30% for the encoder and only 10% for the decoder. Combining this with the two transmission modes mentioned above adds two more columns in Table 1 named “Hardware 30% enc workload, burst transmission” and "Hardware 30% enc workload, continuous transmission ”.

All values in Table 1 are valid for 2-channel stereo mode at an audio sampling frequency of 48 kHz. For each codec a typical stereo bit-rate is used which does not necessarily mean that the audio quality is exactly the same for all the codecs at their typical bit-rate. It

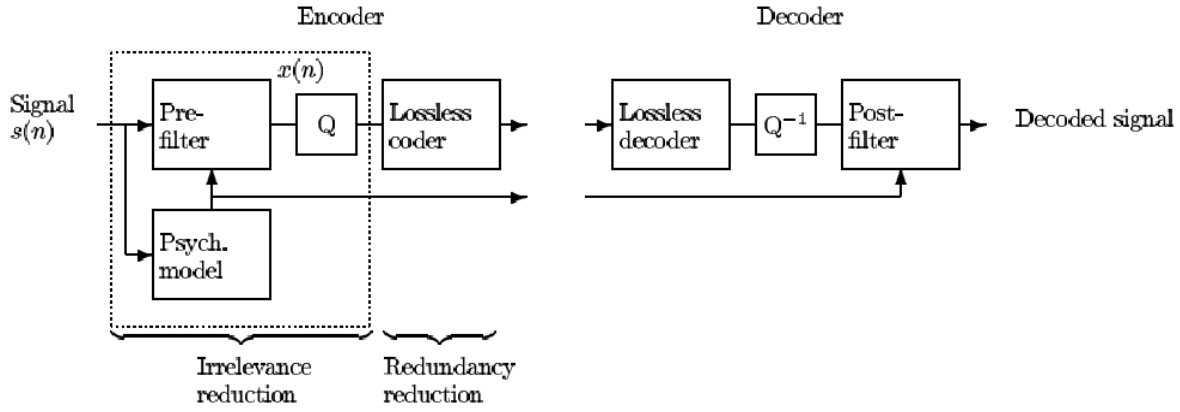

Fig. 6: The audio coding scheme with separated irrelevance and redundancy reduction, using a psycho-acoustic pre- and post-filter and lossless compression.

should also be noted that in the case of continuous bit-stream transmission this favors codecs operating at higher typical bit-rates and puts codecs into disadvantage that have lower typical bit-rates and make use of a bit reservoir.

3.2. Predictive coders

3.2.1. G722.1

G722.1 is a low-bit-rate coder which codes 16 kHz sampled audio with a bandwidth of 50 - 7500 Hz at 24 kbps or 32 kbps. The quality at 32 kbps is comparable to that of G722 SB-ADPCM at 64 kbps. The coding scheme is based on a MDCT with 320 subbands and 620 tap window, hence a system delay of 619 samples.. At 16 kHz sampling rate this results in a frame size of 20 ms and a algorithmic delay of 40 ms.

3.2.2. G722.2

The G722.2 which is also known as ARM-WB codec utilizes the Algebraic Code Excited Linear Prediction (ACELP) technology. As it descends from the speech coding family its audio quality degrades using it for music signals. The range of bit-rates lies between 6.6 and 23.85 kbps. At a block-length of 320 samples, a look ahead of 5ms and a sampling-rate of 16 kHz it comes up with an algorithmic delay of 25ms [10][11].

3.2.3. apt-X

The apt-X codec is more similar to wide band speech-codec G.722 than to other audio coding systems. A four band QMF filter bank is used to split the audio signal into frequency bands of equal bandwidth. The subband signal is ADPCM coded with a fixed bit allocation. The bit-rate ranges from 64 kbps (mono, 16 kHz sampling-rate, 7.6 ms delay) to 576 kbps (stereo, 48 kHz sampling-rate, 1.9 ms delay) [9].

3.2.4. Fraunhofer Ultra Low Delay (ULD)

Predictive coders have traditionally been used in speech coding. An example are well known ADPCM coders. Examples are G.726, G.727, G.722 [13]. They are for bit-rates used for speech coding, in the range of about 16 to 64 kb/s. Their advantage is a very low encoding/decoding delay, important for communications applications. It is about 0.125 to 1.5 ms. The problem for higher quality applications is, that they don't have the desired audio quality. In order to obtain an audio coder with the desired quality and a low delay, we can take advantage of the predictive structure of these coders, which leads to the low delay, and build an audio coder on that basis.

The main problem of this approach for audio coding is that psycho-acoustic models are based on a subband decomposition of the audio signal, hence there is no direct way to apply the output of a psycho-acoustic model to predictive coding.

To solve this problem the ULD coder [14] separates the application of psycho-acoustics (the irrelevance reduction) from the redundancy reduction, so that 2 separate units [15] are obtained. The input of the

psycho-acoustic model still consists of subband signals from an analysis filter bank. But since the irrelevance reduction unit is not constrained by coding efficiency, the number of subbands can be chosen smaller. 128 uniform bands are chosen and it was found to give sufficient frequency and also time resolution for time and frequency masking effects. The output of the psycho-acoustic model is the masking threshold for each subband. The pre-filter is a linear filter which has a structure like a predictor. Its coefficients are continuously chosen such that its magnitude response is the inverse of the masking threshold. It "normalizes" the signal to its masking threshold. The decoder has a post-filter, which reverses this normalization. Hence it has a magnitude response like the masking threshold for the signal. Since the post-filter needs to be the inverse of the pre-filter, side-information for the filter coefficients is necessary.

Quantization is conducted after the pre-filter. Since at that point the signal is normalized to its masking threshold, the quantization error needs to be flat over frequency, and a simple uniform quantizer can be used, in practice just a rounding operation. The delay of this stage is determined by the psycho-acoustic model and its filter bank. In the Fraunhofer implementation the filter bank has 128 subbands with a filter length of 256, leading to a corresponding delay. Since only an analysis filter bank is needed, the delay can be seen as about 128 samples (the peak of the filter bank window). Additionally an interpolation of the masking threshold between blocks is useful, which introduces another 128 samples, totaling 256 samples.

After applying psycho-acoustic effects for the irrelevance reduction in the first stage, it can be followed by a lossless predictive coder, which can be seen as the stage for redundancy reduction. Current lossless audio coders are typically based on block wise forward prediction. The prediction coefficients for a block are transmitted as overhead, and the residuals are Huffman coded and transmitted. This means there is a delay of at least one block size. To obtain a low delay, backward adaptive predictive coding can be used, which is also a standard techniques in low-delay speech coding [16]. Backward adaptive prediction has also been used in previously in audio coding, e.g., in backward adaptive warped lattice algorithm proposed in [17]. However, in these cases backward adaptive prediction was used in LPC filters, and in a lossy scheme, while here it is used in lossless compression after the quantizer.

The backward adaptive prediction is implemented using the normalized least means squares (NLMS) algorithm [18]. Since backward adaptive predictive coding is used in this second stage, there is no delay here. The predictor is followed by an entropy coder. Entropy coders with low delay can be used, for instance arithmetic coding, adaptive Huffman coding, or Golomb coding. Golomb coding has the advantage that it has a low computational complexity and also no delay.

Hence the overall encoding/decoding delay is on the order of 128 samples, plus the possible delays for the interpolation of the pre-filter coefficients, leading to a total of about 256 samples or about 8 ms at 32 kHz sampling.

The range of possible bitrates lies around 24 to 96 kbps/ch. High audio quality can be achieved at 70 kbps/ch.

4. CONCLUSIONS

It can be seen that there is a wide variety of audio coders available, for a wide variety of applications. Traditionally, coders where made for high quality, high delay, and high compression ratio (MP3, MPEG-AAC), or for low delay, low bit-rate, and speech quality (speech coders). In the meantime, other combinations are available, for instance for high quality communications applications. Examples are the AAC Low Delay coder with about 20 ms delay at 48 kHz sampling rate, and compression ratios and quality comparable to or better than MP3; or the Ultra Low Delay Coder, which is based on prediction. It combines about 8 ms delay at 32 kHz sampling with high audio quality and state-of-the-art compression ratios.

5. REFERENCES

[1] G. Schuller, B. Yu, D. Huang, and B. Edler, "Perceptual Audio Coding using Adaptive Pre- and Post-Filters and Lossless Compression", IEEE Transactions on Speech and Audio Processing, September 2002, pp. 379-390

[2] Th. Sporer, K. Brandenburg, and B. Edler. "The use of multirate filter banks for coding of high quality digital audio", In 6th European Signal Processing Conference (EUSIPCO), volume 1, pages 211-214, Amsterdam, June 1992. Elsevier.

[3] E. Allamanche, R. Geiger, J. Herre, Th. Sporer. "MPEG-4 Low Delay Audio Coding Based on the AAC Codec", 107th AES-Convention, Munich 1999. preprint 4929

[4] K. Brandenburg "MP3 and AAC explained", 17th AES conference on High Quality Audio Coding, September 1999

[5] Karlheinz Brandenburg and Marina Bosi. "Overview of MPEG Audio:Current and Future Standards for Low-Bit-Rate Audio Coding", 99th AES convention, Dezember 1995

[6] Frans de Bont, Marc Groenewegen, Werner Oomen, "A High Quality Audio - Coding system at 128 kb/s", 98th AES Convention, Febr. 25-28, 1995

[7] http://qualweb.bluetooth.org

[8] Bernhard Grill, Harald Gernhardt, Michael Härtl, Johannes Hilpert, Manfred Lutzky and Martin Weishart, "MP3 on MPEG-4", 115th AES convention, October 2003

[9] http://www.aptx.com/sitefiles/resources/aptxoverview.pdf "Digital Audio Data Compression Technical Overview"

[10] ITU-T Study Group 16, Liaison to multiple SDOs requesting input for "Media Coding Summary Database" Project, Geneva, 15-25 October 2002

[11] 3GPP TS 26.090 V5.0.0 (2002-06)

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

[13] V.K. Madisetti, D.B. Williams, "DSP handbook", IEEE press 1997

[14] G. Schuller, A. Harma: "Low Delay Audio Compression using Predictive Coding", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, FL, May 13-17, 2002

[15] G. Schuller, B. Yu, D. Huang, and B. Edler, "Perceptual Audio Coding using Adaptive Pre- and Post-Filters and Lossless Compression", IEEE Transactions on Speech and Audio Processing, September 2002, pp. 379-390

[16] J.-H. Chen, R. V. Cox, Y.-C. Lin, N. Jayant, and M. J. Melchner, "A low-delay CELP coder for the CCITT 16 kb/s speech coding standard" IEEE J. Sel Areas in Comm, vol. 10, pp. 830--849, June 1992

[17] A. Harma, U. K. Laine, and M. Karjalainen, "Backward adaptive warped lattice for wideband stereo coding" in Proc. of EUSIPCO '98, Greece, 1998

[18] S. S. Haykin (1999). "Adaptive Filter Theory". Englewood Cliffs, N.J. : Prentice Hall

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

[20] ISO/IEC JTC1/SC29/WG11 (MPEG), International Standard ISO/IEC 13818-7 "Generic Coding of Moving Pictures and Associated Audio: Advanced Audio Coding", 1997

[21] ISO/IEC JTC1/SC29/WG11 (MPEG), International Standard ISO/IEC IS 14496-3: "Coding of Audio-Visual Objects: Audio", 1999

[22] ISO/IEC JTC1/SC29/WG11 MPEG2003/N5711 "Proposed ISO/IEC14496-3 (Audio 3rd Edition)" subpart 4