# AN AUDIO CODEC FOR MULTIPLE GENERATIONS COMPRESSION WITHOUT LOSS OF PERCEPTUAL QUALITY

**FRANK KURTH**

*Department of Computer Science V, University of Bonn, Bonn, Germany*
frank@cs.uni-bonn.de

We describe a generic audio codec allowing for multiple, i.e., cascaded, lossy compression without loss of perceptual quality as compared to the first generation of compressed audio. For this sake we transfer encoding information to all subsequent codecs in a cascade. The supplemental information is embedded in the decoded audio signal without causing degradations. The new method is applicable to a wide range of current audio codecs as documented by our MPEG-1 implementation.

## INTRODUCTION

Low bit rate high quality coding is used in a wide range of nowadays audio applications such as digital audio broadcasting or network conferencing. Although the decoded versions of the compressed data maintain very high sound quality, multiple- or tandem-coding may result in accumulated coding errors resulting from lossy data reduction schemes. Such multiple or tandem coding, leading to the notion of a signal's *generations*, may be decribed as follows. Assuming a coder operation $C$ and a corresponding decoder operation $D$ we shall call, for a given signal $x$, $DCx$ the first generation and, for an integer $n$, $(DC)^n x$ the $n$-th generation of $x$.

In this paper we propose a method to overcome *ageing effects*. More precisely, our ultimate goal is to preserve the *first* generation's perceptual quality. For this sake we use an embedding technique, conceptually similar to the audio mole [1], to transport coding information from one codec in a cascade to subsequent codecs. As compared to [1], our embedding technique is performed in the transform domain. Moreover it is based on psychoacoustic principles which allows the embedding to be performed without causing audible distortions.

The paper is organized as follows. In the first section we perform a detailed analysis of ageing effects. For this sake we look at a general model of a psychoacoustic codec and investigate which of its components may induce artifacts in casacded coding. Ageing effects in a real-world coding application are documented by the results of listening tests as well as measurements of relevant encoding parameters. The second section develops a generic audio codec for cascaded coding without perceptual loss of quality. In the third section, an implementation based on an MPEG-1 Layer II codec is described. The fifth section gives a codec evaluation based on extensive listening tests as well as objective signal similarity measurements. Concluding we point out some applications and future work in this area.

## 1. AGEING EFFECTS IN CASCADED CODING

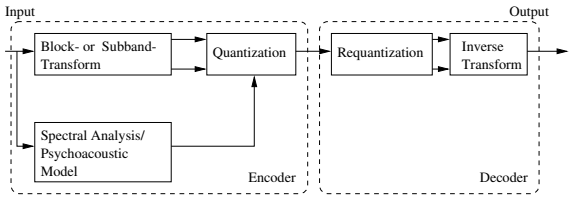


Figure 1: Simplified scheme of psychoacoustic codec.

A general scheme of a psychoacoustic audio codec is given in Fig. 1. In this figure we have omitted optional components such as entropy coding or scalefactor calculation, although those may indirectly influence degeneration effects. Such effects will have to be dealt with in each particular case.

The encoder unit consists of a subband transform $T$ operating in a block by block mode. For a signal block $x$, a synchronous (w.r.t. the subband transform of that block) spectral analysis $Sx$ is used to obtain parameters of a psychoacoustic model $\Psi(x)$. A bit allocation $b(\Psi(x), Tx, x)$ is performed using the psychoacoustic parameters and leading to a choice of quantizers for lossy data reduction. Following quantization, codewords and side information, e.g., quantizers, scalefactors etc., are transmitted to the decoder. The decoder reconstructs subband samples using the side information, e.g., by dequantization, codebook look-up, or inverse scaling, and then carries out a reconstruction subband transform $\tilde{T}$.

Within this framework, sources for signal degeneration are

1. the lossy quantization step,
2. round-off and arithmetic errors due to $T$, $\tilde{T}$, $S$, as well as possibly further calculations,
3. aliasing due to $T$, cancelled by $\tilde{T}$ only in the absence of lossy quantization,

4. NPR (near perfect reconstruction) errors, i.e., if $\tilde{T}$ is only a pseudoinverse of $T$,
5. inaccuracy of the psychoacoustic model $\Psi$,
6. non-suitable time-frequency resolution in spectral analysis,
7. missing translation invariance of $T$ and $\tilde{T}$ (in view of multiple coding).

The quantization error 1. dominates and is likely to be responsible for the greatest part of the degenerations. This error may differ in several magnitudes from all of the other errors. As an extreme example, consider MPEG coding where, depending on the codec configuration, several of the highest subbands of the transformed signal are simply set to zero. Round-off and NPR-errors (2. and 4.) are small as compared to 1., although they must be controlled from codec to codec in a cascade. Those errors will be especially important in conjunction with our embedding technique. As, e.g., Layer III of MPEG-1 audio coding shows, aliasing errors (3.) have to be dealt with carefully. The errors 5. and 6. indirectly contribute to the quantization error.

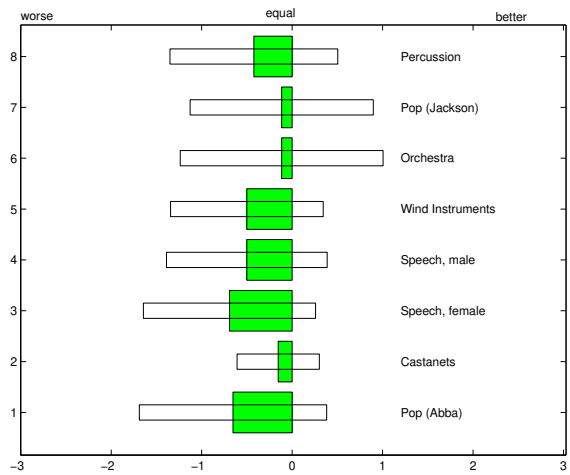


Figure 2: Comparison of first vs third generations. The solid bars give the average rating, the small bars show the variance.

Listening tests are used to investigate signal degeneration with increasing generations. Fig. 2 shows the results of a listening test where 26 listeners compared first and third generations coded with an MPEG-1 Layer II codec at 128 kbps. Negative values indicate that the third generations were rated worse than first generations, whereas positive values give them better ratings. It is obvious from Fig. 2 that the third generations could almost all be distinguished from the first ones and were generally judged to be of worse quality.

The most important results [2] may be summarized as follows:

- Almost all test pieces already show noticeable perceptual changes in the second and third generations.
- Fourth to sixth generations show a clear loss in sound quality. The induced noise in many of the pieces is annoying. Almost all pieces above the eighth generation show a severe perceptual distortion.
- Quiet and very harmonic signal parts are (almost) not distorted due to MPEGs extensive use of scale factors and due to higher SMR ratios induced by tonal signal components.

It becomes clear that lossy coding changes the signals spectral content in a way that does not allow subsequent encoders to perform a suitable psychoacoustic analysis. This leads to a degeneration of coding parameters and, finally, of overall sound quality. We further illustrate this

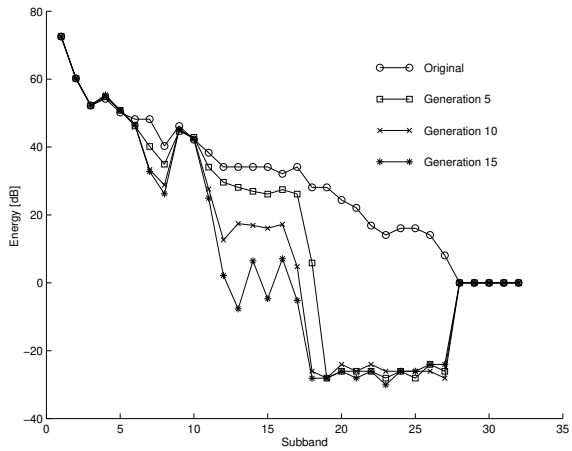


Figure 3: Degeneration of subband energies in MPEG-1 coding. For a fixed frame, the figure shows the energy in each of the subbands. The four lines correspond to the energies of the original signal and the fifth, tenth, and fifteenth generation respectively.

by an example documenting the change of the spectral content of a signal for increasing generations. Fig. 3 shows the subband energies for a fixed frame in the case of an MPEG-1 encoded guitar piece. The different lines represent the energies for different generations. One observes the decrease of energy in subbands 6-9 and 11-17. The higher subbands are attenuated as an effect of the MPEG encoder not allocating any bits to them. In Fig. 4 the energy of the scalefactor bands is given for several frames of a piece containing a strummed electric guitar

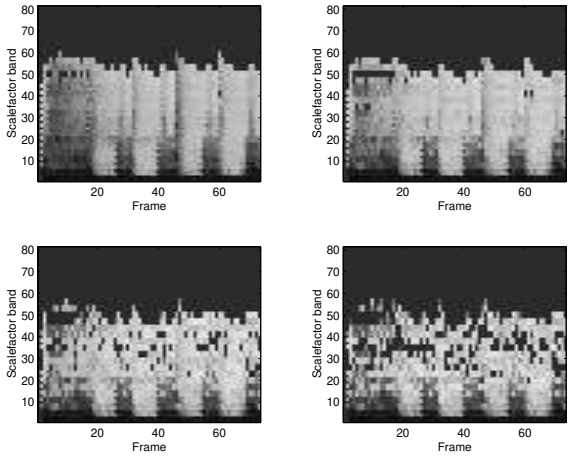


Figure 4: Degeneration of subband energies in MPEG-1 coding. For successive frames, the figure shows the energy in each of the scalefactor bands as an intensity plot. The four subplots correspond to the energies of the first, fifth, tenth, and fifteenth generation respectively.

(the rhythm of strumming may be observed as the vertical structure in each of the plots). Increasing generations are plotted in a zig-zag scheme. The signal's degeneration is obvious from the loss of structure. Black regions indicate that scalefactor bands are set to zero during bit allocation.

We briefly mention the effect of missing translation invariance of the coding operation. As already observed in [3], a crucial point in cascaded coding is *synchronicity*. Taking the 32-band multirate MPEG-1 filter bank as an example, a subsequent encoder will not obtain the same subband samples as the preceeding decoder unless the input signal is "in phase" with the 32-band filter bank. This is, the filter bank is only translation invariant w.r.t. signal shifts of 32 samples. On a frame by frame basis, things are even worse, since in order to obtain the same content in a *framewise* manner, the input to the second coder has to be frame-aligned (e.g. within 1152 samples in the MPEG-1 Layer II case). Hence, we may also have various types of generation effects, depending on eventual signal shifts prior to subsequent encoding. Listening tests in our MPEG-1 Layer II framework show remarkable differeces in the *type* of degeneration depending on whether the signal was

1. fed into subsequent codecs *without* any synchronization, or
2. fully synchronized to the original input and then fed into the codec.

Whereas the unsynchronized signals tend to produce audible artifacts and noise-like components, the synchronized versions tend to attenuate several frequency bands.

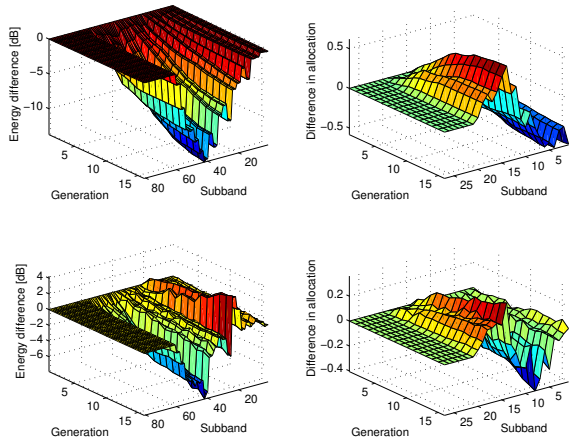


Figure 5: Development of mean subband energy (left) and mean bit allocation (right) as a difference signal of (increasing) generations and the first generation. The top plots show those difference signals for the case of synchronization between each of the codec cascades, the bottom plots for the case of absence of such a synchronization.

It depends on the signal, which of the degenerations are percieved to be more severe. Fig. 5 shows four difference plots (first generation versus higher generations) for the scalefactor bands mean energies and mean difference in bit allocation, where the mean ranges over 200 frames of a relatively stationary signal. Clearly, in the case of synchronization between subsequent coding (top left plot), the signal energy is attenuated, whereas absence of synchronization (bottom left plot) shows no such tendency resp. distributes the energy across the subbands.

It is remarkable that all formerly proposed methods for cascaded audio coding [1, 3] take into account a synchronization mechanism to overcome the mentioned distortion effects. Yet it should be clear that such a mechanism alone is not sufficient to overcome degeneration.

## 2. A CODEC PREVENTING AGEING EFFECTS

The idea behind our new coding method is to reuse the first generation's encoding parameters in all of the subsequent codecs in a cascade. This allows us to indirectly use the psychoacoustic analysis of the first encoder in all of the following encoders. Since the encoding parameters are not present at the point of decoding, we use the *decoding* parameters, i.e., sideinfos. This is reasonable for in most cases we may derive the encoding parameters from them.

While it first seems trivial that one may reuse the encoding information of a certain encoding operation for subsequent encodings, it is not at all clear

- that one may deduce all encoding information from

the *decoding information* only,

- that it is possible, using this information, to perfectly reproduce the compressed bitstream of the first generation (requantization and scaling could not be injective),
- how to convey this information from one codec to another.

While the first two issues depend on the particular codec, the last one poses a general problem, for it is in general undesirable to create a secondary bitstream or file format. Such a secondary bitstream would cause additional data overhead, would require a new format definition, and, most important, would make it very difficult to use the proposed technique in conjunction with standard media not supporting the new format. For this sake, the secondary information for subsequent encoding should be *embedded into* the decoded PCM data. The use of such a *steganographic* technique [4] in this framework was also independently developed by Fletcher [1]. However, Fletcher's embedding method significantly differs from the method presented in this paper leading to rather different codec structures.

### 2.1. Psychoacoustic Embedding

The embedding of secondary data into some target data stream is known as steganography and has been studied to a considerable extent [4]. Our demands on a steganographic process are the absence of any perceptual degradation in the target signal, a high embedding capacity on a frame by frame basis, and a computationally efficient embedding procedure. Furthermore, efficient detection and extraction of the embedded data must be possible.
For a straightforward embedding approach consider a sample $b = \sum_{j=0}^{n-1} b_j 2^j$ also written in its binary representation $b = (b_{n-1} \dots b_0)$, $b_j \in \{0, 1\}$. For $k < n$ and an embedding word $z = (z_{k-1} \dots z_0)$ we define the $k$-bit (direct) *embedding* by

$$E_1 : (b_{n-1} \dots b_0), (z_{k-1} \dots z_0) \mapsto (b_{n-1}..b_k z_{k-1}..z_0). \quad (1)$$

This kind of embedding destroys the $k$ least significant bits of the data word $b$. Application of this technique to a time signal already causes a noticeable noise level for small $k$. In the audio mole proposal [1], this embedding technique is used with the least significant bit(s). Depending on the bit resolution, the possibility of small audible distortions are reported [5]. As an alternative to overwriting the least significant bit, the authors propose to change it according to a parity/non parity decision, which causes a more random signal change.
A refinement of the time-domain embedding technique uses an invertible linear transform $T$ prior to embedding. Embedding now becomes a map

$$(x, z) \mapsto T^{-1} E_1(Tx, z). \quad (2)$$

From a steganographic point of view, a significant advantage of this method is that the embedding is not as easy to detect as the above without knowledge of $T$. Yet it is not guaranteed that the reconstructed signal is still of perceptually transparent quality.
From a psychoacoustic coding point of view it seems to be adviseable to perform a kind of selective embedding. Embedding is performed prior to applying the reconstruction transform $\tilde{T}$. The embedding positions and -widths (e.g., $k$ above) are given by the psychoacoustic parameters or, in view of embedding in the decoder unit, by the decoding parameters. Intuitively, the possible choice of reconstruction levels for a given quantizer $Q$ with requantizer $\tilde{Q}$ is increased with the amount of data reduction, i.e., decreasing number of reconstruction levels. The embedding thus becomes a map

$$(x, z) \mapsto T^{-1} \tilde{E}_1(\tilde{Q}(QTx), z, Q), \quad (3)$$

where the embedding width of $\tilde{E}_1$ depends on $Q$. This technique is used in the proposed embedding codec presented in the next paragraph.

### 2.2. Codec Model

We first give an overview of the proposed codecs functionality. Fig. 6 shows the generic codec scheme. To

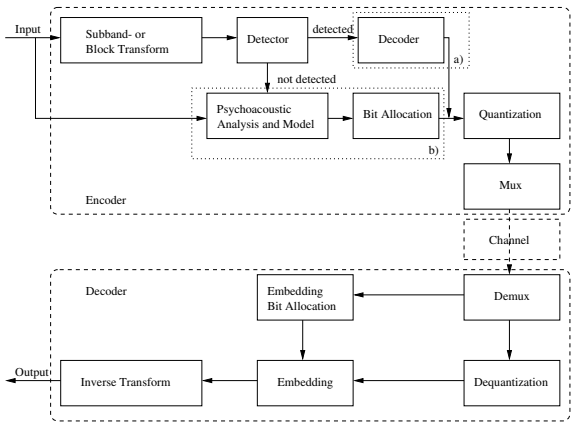


Figure 6: Codec for direct embedding.

describe the functionality, we start with the *decoder*. The decoder extracts codewords and side information for one data block from the bitstream and proceeds by dequantization. All parameters needed by the next encoder to reproduce the compressed bitstream are assembled in a bit buffer. Then, using the quantization information, a bit allocation algorithm decides which subbands will be used for embedding, and how many bits will be embedded in

each of them. If enough bits could be allocated, the decoder proceeds by embedding all data from the bit buffer. To allow extraction of the embedded data in subsequent encoding steps, certain *markers*, i.e., bit combinations, are used to indicate embedding. Afterwards, the inverse transform reconstructs the time signal block.
The *encoder* proceeds by transforming the signal blocks producing subband signals. For this sake, we assume that encoding is *sychronous* to the decoding operation w.r.t. the processed signal blocks. The *detector* tries to detect the markers created by the decoder. If markers are found, a decoder extracts the embedded information from the subbands. Using the embedded information, the bit allocation and further encoder parameters can be set accordingly. If, however, no markers are found or the embedded information is detected to be corrupt, the encoder follows the standard encoding procedure using the psychoacoustic model. Quantization and channel multiplexing conclude the encoding.
The single modules deserve a more detailed treatment, where we shall again start with the decoder:

- To obtain the embedding capacity, i.e., the number of bits per sample which we may use to store our secondary data stream, we divide the subband samples (or transform blocks in transform coding) in groups of samples sharing the same quantizers. Those groups are called *embedding blocks*. For each embedding block we calculate the number of bits available for embedding, which depends on the quantizer's granularity. If, e.g., the quantizer reduces an $n$-bit sample to a $k < n$ bit representation, we may use at most $n-k$ bits for embedding in this sample. Note that this direct correspondence is not valid for all kinds of quantizers and more elaborate conversion rules have to be used by non-uniform quantizers. It is also important to note that lossless coding operations might eventually limit the embedding capacity, as it will show up in the case of MPEG's scalefactors.

- All information to be embedded is collected in an embedding bitstream or bit buffer. The type of information depends on the chosen codec, examples are bit allocation, quantizers, scalefactors, codebooks, bitrates etc. Sometimes it will be useful to exploit the compressed bitstream's structure, since most of the latter information is already stored in this bitstream in a very compact form.

- If the total amount of embedding capacity suffices to transmit the desired coding information, we have to select the embedding blocks where we want to store the embedding bit stream. Furthermore, we have to determine an embedding bit width per embedding block. The decision about those parameters is termed *embedding bit allocation*. The embedding bit allocation, although consuming only a fraction of computational resources as compared to the encoder's bit allocation, is the most complex part of the proposed codec extension.

  Since the proposed embedding technique theoretically may yield a bigger reconstruction error than normally induced by the utilized quantizers, a conservative allocation method should be used. We propose a greedy bit allocation algorithm with embedding blocks sorted in descending order w.r.t. their embedding capacities. To each of the embedding blocks we assign a certain bit budget, obeying a safety margin accounting for the possible increased reconstruction error. If the greedy bit allocation loop does not yield enough embedding capacity, the bit budget is increased as long as no further increase is possible. Embedding is only performed if the allocation procedure succeeds.

- To facilitate the detection of the embedded information, the kind and position of the embedding have to be signaled to the encoder resp. its detector unit. This amounts to conveying the embedding blocks and embedding bit widths chosen by the embedding bit allocation. Several techniques are possible, e.g., the use of a descriptor embedding block containing all of this "logistic" information. We propose a block-by-block solution where each used embedding block contains a separate marker indicating the corresponding bit width. We only have to ensure that no "wrong detections" (i.e., the embedding bit width is not detected correctly) occur.

- Since floating point operations commonly causes arithmetic errors, a forward error correction (FEC) has eventually to be applied to the embedded information. Those errors may accumulate with NPR errors as discussed above. A CRC checksum can be used to decide whether the information extracted from a certain embedding block is valid or not.

- The first task of the encoder is to synchronize the PCM bitstream w.r.t. the frame boundaries used by a previous codec. For this purpose, again several techniques may be utilized. In the case of a filter bank transform, a (fast) search for markers on certain predetermined (frequently used) subbands may be used. Once the synchronization is done, the encoder proceeds on a frame by frame basis and no further synchronization is necessary.

- The detector tries to find the embedded markers, deduce the embedding blocks embedding widths and check the embedding blocks integrity.

- Hybrid coding provides a mechanism against corrupt embedded data or frames where no embedding was possible. In a hybrid framework, for each frame it is possible to choose between
  - usage of all embedded coding information if this information could be extracted,
  - partial usage of embedded information, e.g. bit allocation only,
  - discarding of all of the embedded information and use of the standard encoding procedure.

**Input:** Codewords $q(s)$, decoding parameters $D$, embedding function $E_1$

1. Requantize $q(s) \mapsto \tilde{s} =: (\tilde{s}_1, \ldots, \tilde{s}_n)$.
2. Determine embedding positions and widths,
$$(\tilde{s}, D) \mapsto (i, t) =: ((i_1, \ldots, i_m), (t_1, \ldots, t_m)),$$
as well as a suitable marker $M$, such that $1 \leq i_k < i_{k+1} \leq n$, and a suitable binary representation $(D, M) \mapsto \text{bin}(D, M)$, consisting of $|\text{bin}(D, M)| = \sum_{j=1}^{m} t_j$ bits.
Here, $t_j$ denotes the embedding width in bits of position $j$.
3. Create a partition $\text{bin}(D, M) =: e_1 \cdots e_m$, where $e_j \in \{0, 1\}^{t_j}$.
4. For $1 \leq j \leq m$ embed into $\tilde{s}$ using $t_j$-bit embedding:
$$E_1 : (\tilde{s}_{i_j}, e_j) \mapsto \tilde{\tilde{s}}_{i_j}.$$
5. Let $\tilde{\tilde{s}}_k := \tilde{s}_k$ for all $k$, where $k \neq i_j$ for $1 \leq j < m$.
6. Reconstruct $\tilde{\tilde{s}} \mapsto T^{-1}\tilde{\tilde{s}}$.

**Output:** Signal block $T^{-1}\tilde{\tilde{s}}$.

Figure 7: *Pseudo code version of the decoding algorithm.*

A pseudo code version of the decoding algorithm is given in Fig. 7. We assume that codewords $q(s)$ as well as decoding parameters $D$ are given as an input. In this example, $q$ denotes the encoders quantizer function. For sake of simplicity we only use fixed quantizers as well as a global marker $M$. Note that we separate encoding and decoding parameters to clarify the coding steps. The embedding bit allocation is summarized in 2. In 3., the embedding bit stream is partitioned according to the bit allocation. Finally, embedding is performed in 4. prior to the reconstruction transform.

**Input:** Signal block $x \in \mathbb{R}^n$, transform $T$ psychoacoustic model $\Psi$, bit allocation $b$, detector function $\mathcal{D}$ on $\mathbb{R}^n$, $\mathcal{M} \subset \text{Im}(\mathcal{D})$ denotes the set of all valid markers

1. Calculate $s := Tx$.
2. Detect $\mathcal{D}(s) =: M$.
3. If $M \notin \mathcal{M}$
   (a) Calculate $\Psi(x)$. Let $s' := s$.
   (b) Calculate $b(\Psi(x), s, x)$ and from this encoding parameters $E = E(x)$.
4. else
   (a) Decode $s \mapsto \tilde{q}(s) =: (s', D)$.
   (b) Calculate encoding parameters $D \mapsto E$.
5. Quantize $s'$ using $E$: $s' \mapsto q(s')$.
6. Determine decoding parameters $E \mapsto D$.

**Output:** codewords $q(s')$, decoding parameters $D$.

Figure 8: *Pseudo code version of the encoding algorithm.*

We summarize the encoding algorithm in Fig. 8. In the pseudo code, the function symbols for the bit allocation and psychoacoustic model are chosen as in the initial discussions. The set of admissible or valid markers used to indicate embedding is denoted by $\mathcal{M}$. The detector function is named $\mathcal{D}$, quantizer and dequantizer are $q$ and $\tilde{q}$, respectively. The heart of the encoding algorithm lies in the decision whether to use the embedded information 4. or not 3. and use the standard procedure instead.
It is important to note that several design decisions including the choice of embedding blocks and embedding bit allocation heavily depend on the underlying codec. Such decisions will be treated in the next section.

## 3. MPEG-1 LAYER II IMPLEMENTATION

To implement a codec preventing ageing effects we chose MPEG-1 Layer II [6] as a basis. The well-know coding scheme is depicted in Fig. 9. Although, strictly speaking, only the bitstream syntax and decoding are standardized, in what follows we shall also talk of *the* encoder. Relating the Layer II codec to our generic scheme, we have a 32-band multirate filterbank as a transform and a psychoacoustic model based on a windowed Fourier spectrum. The psychoacoustic model yields a signal-to-mask ratio used in the bit allocation process. To reduce amplitude redundancy, blocks of subband samples are assigned common scalefactors and scaled accordingly. The scaled values are linearly quantized using a variable quantizer step size according to the bit allocation. Codewords and side information are stored in the MPEG bitstream and trans-

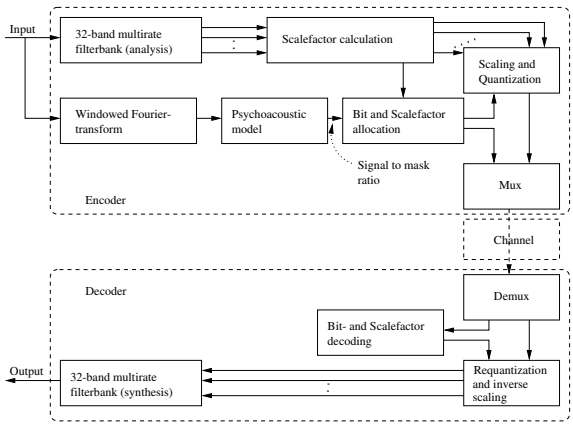


Figure 9: MPEG-1 Layer II codec.

ported to the decoder.
To extend the MPEG Layer II codec to the proposed novel codec scheme, we consider

- the choice of suitable embedding blocks,
- the side information to be embedded and the embedding bit buffer,
- the determination of the available embedding bit widths,
- the embedding bit allocation,
- error correction mechanisms, and
- the choice of suitable markers.

### 3.1. Embedding blocks

MPEG-coding works on a frame by frame basis. In Layer II, blocks of 1152 samples per channel are filtered, yielding 36 samples in each of the 32 subbands. Scalefactors are assigned to blocks of 12, 24 or 36 samples within one subband, depending on the magnitudes of the samples. Since scalefactors implicitly change the quantization resolution, the embedding capacity depends on the particular scalefactor assignment/combination. For simplicity, we assume the worst-case of 12-sample scalefactor blocks and choose those blocks as embedding blocks. Hence there are potentially three embedding blocks per subband and frame. The following discussion assumes a fixed given MPEG frame.

### 3.2. Embedded information and bit buffer

Besides global information such as the number of channels, bitrate or joint stereo coding modes, we have to transmit/embed the following frame-related information:

- Bit allocation: depending on the target bit rate a variable number of subbands is allowed for bit allocation. In MPEG, 2-4 bits are used for each subband's allocation information.
- Scalefactor selection: each subband with a positive number of allocated bits is assigned 1-3 scale factors, depending on the subband samples magnitudes. The select-information consists of 2 bits conveying the utilized scalefactor pattern.
- Scalefactors: for each of the 1-3 scalefactors, 6 bits are used.

First, one might think that the transmission of scalefactors could be obsolete for they only represent a lossless coding step. Yet since dequantization eventually significantly alters a sample's magnitude, a subsequent encoder might calculate different scalefactors, again implicitly resulting in a different quantizer step size. Thus we transmit both scalefactors and select information. Listening tests for the case where we only transmitted bit allocation information show a significant amount of degeneration due to the afore mentioned effects of scalefactor changes.
To store those parameters, we implemented a bit buffering mechanism consisting of a buffer with read/write access to $n \times k$ bit blocks. The coding parameters are sequentially fed into the buffer. During embedding, the buffer is read in $12 \times k$ bit blocks, corresponding to the embedding block size $n$ and the embedding bit width $k$. In the extractor, the procedure is reversed.

### 3.3. Embedding bit width

For each scaling/embedding block $(k, i)$ (where $k$ denotes the subband number and $1 \leq i \leq 3$ one of the three scalefactors) we determine the maximum embedding bit width $V_{k,i}$ from the corresponding subbands quantizer resolution $A_k$ (in bits) and the assigned scalefactor $s_{k,i}$, $1 \leq i \leq 3$. We have to account for the scalefactors, since they implicitly increase the bit resolution: scaling by a factor $2^{-k}$ prior to quantization allows for an increase in bit resolution by $k$ bits as compared to quantization without scaling. We do not consider subbands with zero bits allocated ($A_k = 0$). In this case we define $V_{k,i} := 0$ for all $i$. In the case that less than three scaling factors are assigned to a specific subband, we determine the $s_{k,i}$ according to the scalefactor pattern.
If the scalefactor $s_{k,i}$ of scalefactor band $(k, i)$ is given by $s_{k,i} = s_n := \sqrt[3]{2}^{(3-n)}$ for $n \in [3 : 62]$, we let

$$V_{k,i} := 16 - \min(16, \lceil -\log_2 s_{k,i} \rceil + A_k).$$

Since $\lceil -\log_2 s_{k,i} \rceil = \lceil -\log_2 \sqrt[3]{2}^{(3-n)} \rceil = \lceil \frac{n}{3} \rceil - 1$ we obtain

$$V_{k,i} = 16 - \min(16, \left\lceil \frac{n}{3} \right\rceil - 1 + A_k).$$

For $n \in [0:2]$, scaling causes the loss of at most one bit of precision. In this case $V_{k,i} := 16 - \min(16, A_k - 1)$. The calculation of $V_{k,i}$ may be motivated as follows: starting from an initial precision of 16 bits, we obtain the MPEG reconstruction precision as the sum of the allocation precision $A_k$ and the precision gained by using scalefactors. The difference between those two quantities and the initial resolution gives the maximum embedding bit width. In case of a precision exceeding 16 bits (which is possible in MPEG), we let $V_{k,i} := 0$.

### 3.4. Embedding bit allocation

For the embedding bit allocation we use a greedy algorithm as already sketched above. We shall only give an overview of the algorithm:

- Sort the embedding blocks in order of decreasing capacities.
- For each block $(k, i)$, assign an initial embedding bit width $e_{k,i} \leq V_{k,i}$.
- Main allocation loop: allocate embedding blocks in the given descending order. Add capacity due to $e_{k,i}$ to counter for embedded bit size. Note that at this point, we have to take care of FEC bits and markers too.
- End the allocation if enough bits could be allocated.
- End the allocation if no change in the allocated bit size occurred as compared to the last loop.
- Increase each $e_{k,i}$ by one bit provided $e_{k,i} \leq V_{k,i}$, then restart the main allocation procedure.

If the embedding bit allocation fails, the decoder does not embed anything for that particular frame.

### 3.5. Error correction and markers

Since the MPEG filterbank does not yield perfect reconstruction, we cannot rely on the encoder's subband samples being identical to those produced by our embedding procedure. There is a certain reconstruction error inherent in the filter bank which accumulates with arithmetic errors possibly introduced by floating point operations. Thus, prior to embedding we perform an arithmetic FEC, mapping the word to be embedded, $x$, to $Ax + B =: y(x)$ for suitable positive integers $A$ and $B$. Then, $y(x)$ is the word to be embedded. As an example, consider $x$ to be an integer, then, choosing $A := 8$, $B := 3$, we may recover $x$ from $y(x) = 8x + 3 + e$ for $e \in [-3, 4]$. In this case, the coding overhead for the arithmetic code would be 3 bits.
Finally, we consider the markers indicating the embedding positions and bit widths. In our implementation we chose variable width markers, embedded at the beginning of our embedding blocks. The marker's width depends on the actual embedding bit width. To prevent false detections in the encoder, we perform an analysis-by-synthesis examination of *all* possible embedding blocks and, if necessary, adapt their content accordingly. Fig. 10 illustrates the concept of an embedding block and gives a rough impression of the embedding positions. For sake of simplicity, we used a 16 bit PCM representation as a reference.

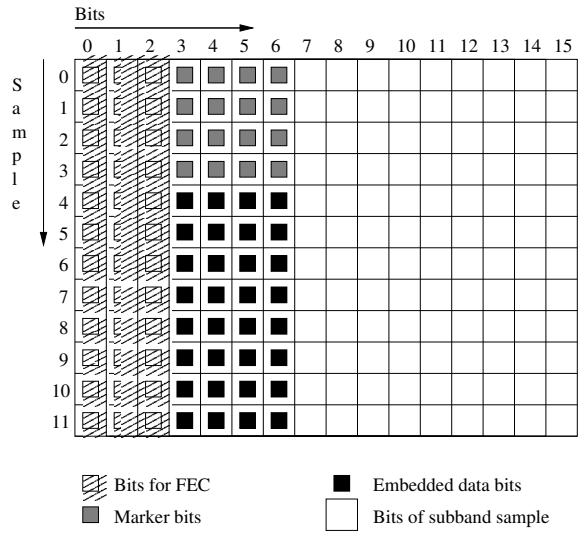


Figure 10: Scheme of an embedding block and the embedding positions. The least significant bits are depicted on the l.h.s.

## 4. CODEC EVALUATION

### 4.1. Test settings

The proposed codec was evaluated on a broad variety of audio pieces, mostly choosen from the widely used EBU test material [7]. We tested our codec on music as well as with male and female speakers. In our test settings, the source material was given as 16 bit PCM stereo with a sampling rate of 44.1 kHz. We used stereo bit rates of 128 to 192 kbps for compression.

### 4.2. Listening tests

In what follows, we shall concentrate on the listening test results for a bitrate of 128 kbps. For our tests we recruited 26 test listeners chosen from among the members of the audio signal processing group and other computer science students at Bonn University.
In this section, the term *standard codec* refers to a non-modified MPEG-1 Layer II codec implementation. In summary we conducted the following tests:

1. Comparison of 5th generations of the standard codec and 5th generations of the proposed codec (rela-

tive).

2. Comparison of first generations and 25th generations of the proposed codec (relative).
3. Comparison of 5th generations using the standard codec and 25th generations of the proposed codec (relative).
4. Comparison of 3rd generations using the standard codec and 5th generations of the proposed codec (relative).
5. Comparison of 10th generations of the standard codec and 25th generations of the proposed codec (relative).
6. Absolute ratings of
   - First generations,
   - 5th generations of the standard codec (without synchronization),
   - 5th generations of the standard codec (including synchronization),
   - 5th generations of the proposed codec, and
   - 15th generations of the proposed codec.

In those tests absolute ratings were given according to the five point MOS impairment scale, while relative ratings were given w.r.t. an integer scale of $[-3, +3]$, indicating if a second stimulus is judged to be of better ($\{+1, +2, +3\}$), worse ($\{-3, -2, -1\}$) or same ($0$) quality as compared to a first stimulus. We shall give an overview of the test results.

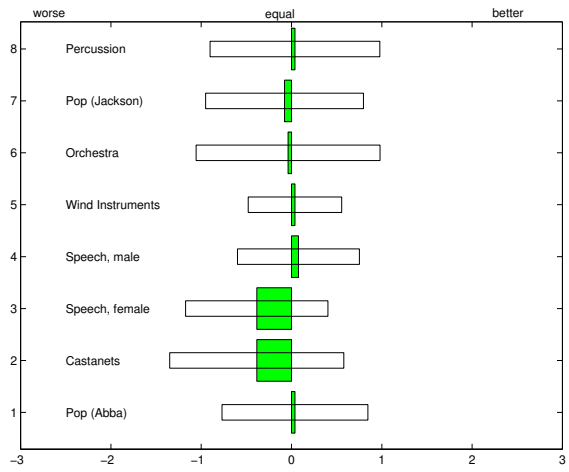


Figure 11: Comparison of first generations and 25th generations using the proposed codec. For most of the pieces, both versions could not be distinguished.

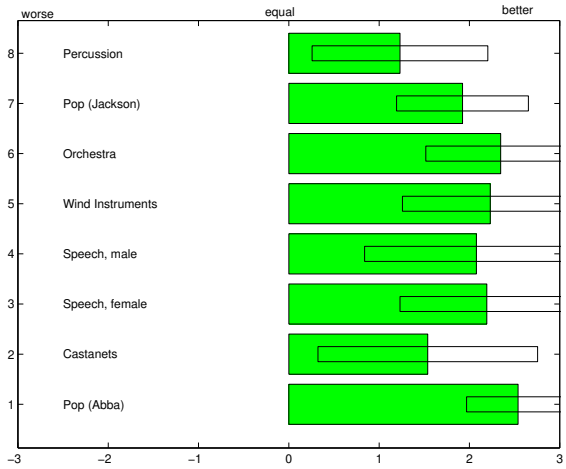


Figure 12: Comparison of 10th generations using the standard codec vs 25th generations using the proposed codec. Positive values give the proposed codec a better rating.

In Fig. 11, the results of the comparison of first generations and 25th generations using the proposed codec are given. It is obvious that for almost all of the pieces, both versions could not be distinguished.
The decrease in quality caused by the standard codec was already reported in the first section. Fig. 12 gives the results for the "extreme" situation of a comparison of 10th generations using the standard codec and 25th generations using the proposed codec. Clearly, the proposed codec is considered to yield much better results. The results of the other tests are consistent with the ones reported here. One problem prevalent in our test settings was the loss of quality in the first generations introduced by the Layer II codec in case of one or two pieces. In those cases, some of the listeners were unsure about which of a piece's versions should be rated worse.
We summarize the trends of our listening tests:

- For six of the test pieces, first generations could not be distinguished from 25th generations generated by the proposed codec.
- For the other pieces, the quality of higher (about 25th) generations are generally judged to be comparable to or better than the 3rd generations produced by the standard codec.
- Tests with high generations (about 50th) show that the signal quality stays stable.

It is important to comment on segments where no embedding is possible. As will be shown in the next subsection, embedding is indeed not possible for each frame. Yet it

turns out that this is not crucial for our MPEG implementation. More precisely, the frames with no possibility of embedding turn out to correspond to quiet signal passages with tonal content. Since

1. the use of scalefactors implies an increased accuracy within those segments and
2. tonal components are quantized in a conservative way (higher signal-to-mask ratio assumed),

there is a natural "workaround" to this problem. Our listening tests confirm this reasoning.

### 4.3. Objective measurements

To support the above results, we give some objective measurements concerning the signal changes with increasing generations. Furthermore, we consider possible (or maximum) embedding capacities being of interest for other steganographic applications.

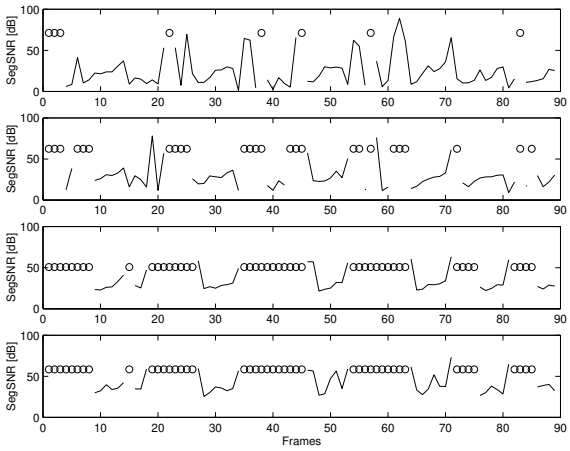


Figure 13: Each of the four parts of the plot each shows (from top to bottom) the segmental SNRs between 1st and 3rd generations, 3rd and 5th generations, 5th and 10th generations, as well as 10th and 25th generations. Small circles at about 60 dB indicate that the corresponding segments are identical

In Fig. 13 the frame by frame SegSNR between various generations of the castagnet piece is given. This is, the segment size is 1152, in synchronicity with the MPEG frames. While the SegSNR is plotted as a solid line, segments where the signal content does not change from generation to generation are indicated by small circles at about 60 dB. It can be observed that with increasing generations (from the top to the bottom plot) there are more and more segments where the signal does not change anymore. This is perfectly in accordance with the observations of the listening test that the signal quality starts to be "stable" starting from about 3rd to 5th generations. In other words, we may say that the proposed codec tends

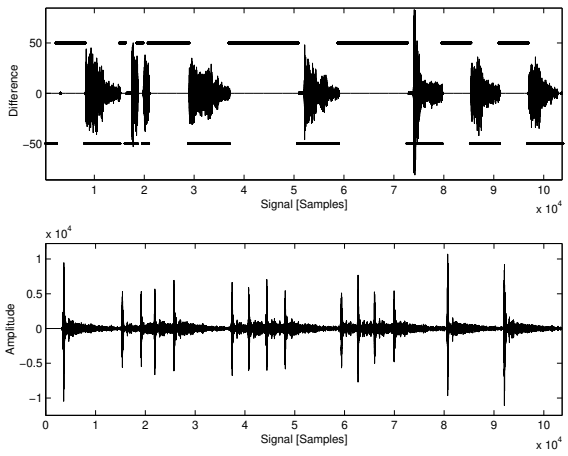


Figure 14: The plot shows the waveform of the castagnet piece (bottom). Above, the difference signal between 5th and 10th generations is given. The upper horizontal bars indicate positions where embedding could be performed while the lower bars indicate segments where no embedding was performed. It is obvious that the difference signal vanishes at the embedding positions (Note that the magnitude of the difference signal is much lower than that of the signal.).

to map the signal partially to some fixed point signal. Comparing the embedding segments to those fixed point positions within the signal, we oberve a match, as depicted in Fig. 14. The lower part of the plot shows the waveform of the castagnet piece. The difference signal between 5th and 10th generations is given in the upper part. The upper horizontal bars indicate positions where embedding could be performed while the lower bars indicate segments where no embedding was performed. It is obvious that the difference signal vanishes at the embedding positions. Measurements with different test pieces and other measures ($\ell^1$-distance, relative segmental SNR) show similar results.

In view of some natural extensions of the proposed codecs as well as other steganographic applications, we examine the embedding capacities obtainable by our apporach. In case of the castagnet piece, Fig. 15 shows the various bit demands and capacities for each frame. The solid line gives the bit demand for the embedding of the coding parameters and the dashed line shows the total embedding capacity. The dotted solid line gives the net embedding capacity, i.e., the total embedding capacity minus the capacity needed for the error correction and the markers. Embedding is only performed for frames where the solid line is plotted below the dotted solid line. One clearly observes the overhead produced by the markers and error correction. Furthermore, some signal parts allow a very huge amount of embedding (about 4000-6000 bits per

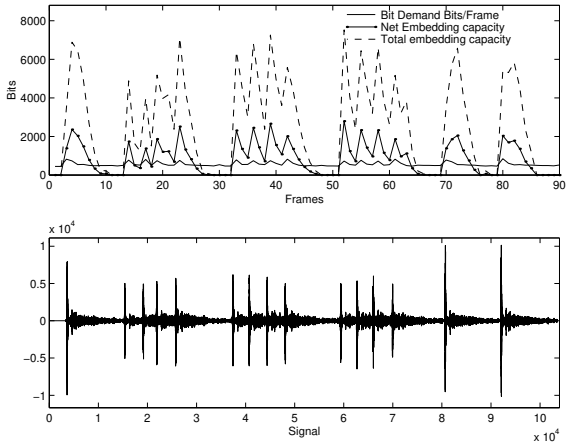


Figure 15: Embedding capacities and demands for the castagnet piece. The solid line gives the bit demand for embedding of the coding parameters (per frame), the dashed line shows the total embedding capacity, and the dotted solid line gives the net embedding capacity.

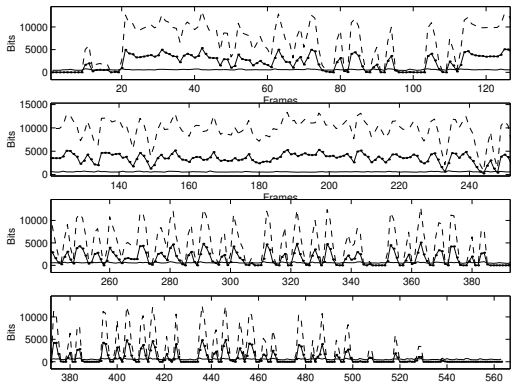


Figure 16: Embedding capacities and demands for a longer segment of a piece of pop music.

frame) which is really a considerable quantity. Note that this is nothing really exceptional, as illustrated in Fig. 16. The plot is analogous to the upper part of Fig. 15 except that a considerably longer signal piece is shown. For this piece of pop music, there are some parts with net embedding capacities of about 4000 bits per frame, in this case corresponding to a total of 10000 bits per frame.

## 5. CONCLUSIONS AND FUTURE WORK

We have presented a novel audio codec scheme suitable for multiple generations audio compression without loss of perceptual quality. As an implementation of the new concept we decribed an MPEG-1 Layer II based codec. A thorough codec evaluation including extensive listening tests and objective measurements shows the proper functionality of the proposed codec. The codec concept is generic and may be combined with a wide range of todays audio codecs (and, naturally, video/image codecs). Further research describing an improved second prototype based on popular transform coding techniques is reported elsewhere.

We also considered steganographic applications. The possible huge embedding capacities as illustrated in Fig. 16 show the flexibility of our embedding approach in view of other applications. Among those is one of our current projects which is concerned with the synchronous integration of textual or score information into the decoded PCM bitstream.

Naturally, there is much room for further improvements of our prototypic implementation including the reduction of FEC/marker overhead using a technique similar to the MPEG-1 Layer III bit reservoir, or introduction of robustness w.r.t. transmission errors. The application of the introduced technique to heterogeneous codec cascades is also of great interest for further studies. Recent developments on a VQ embedding framework allowing for codecs which, under some mild conditions, may be *proven* to keep the perceptual quality of the first generation signal, will be reported elsewhere.

## 6. ACKNOWLEDGEMENT

The author would like to express his gratitude to Michael Clausen for supervising this work and to Meinard Müller for giving several valuable comments. The research activities have been partially supported by Deutsche Forschungsgemeinschaft under grant CL 64/3-1, CL 64/3-2, and CL 64/3-3.

## REFERENCES

[1] John Fletcher, "Iso/mpeg layer 2 - optimum reencoding of decoded audio using a mole signal," in *Proc. 104th AES Convention, Amsterdam*, 1998.

[2] Frank Kurth, *Vermeidung von Generationseffekten in der Audiocodierung (in german)*, Ph.D. thesis, Institut für Informatik V, Universität Bonn, 1999.

[3] Michael Keyhl, Harald Popp, Ernst Eberlein, Karl-Heinz Brandenburg, Heinz Gerhäuser, and Christian Schmidmer, "Verfahren zum kaskadierten Codieren und Decodieren von Audiodaten," 1994, Patentschrift beim Deutschen Patentamt DE 4405659 C1.

[4] Ross J. Anderson and Fabien A.P. Petitcolas, "On the limits of steganography," *IEEE Journal on selected areas in communications*, vol. 16, no. 4, pp. 474–482, May 1998.

[5] A. J. Mason, A. K. McParland, and N. J. Yeadon, "The ATLANTIC audio demonstration equipment," in *106th AES Convention, Munich, Germany*, 1999.

[6] ISO/IEC 11172-3, "Coding of moving pictures and associated audio for digital storage media at up to 1.5 mbits/s- audio part," 1992, International Standard.

[7] Technical Centre of the European Broadcasting Union, "Sound Quality Assessment Material Recodings for Subjective Tests," 1988.