Bit Rates in Audio Source Coding

Raymond N. J. Veldhuis, *Member, IEEE*

***Abstract*—Waveform coding of audio signals at low bit rates generally results in coding errors. In high-quality applications these must remain inaudible. The bit rate required to code audio signals without audible errors depends on both the signal's power spectral density function and masking properties of the human ear. It is shown how rate distortion theory and psychoacoustic models of hearing can be used to compute lower bounds to the bit rate of audio signals with inaudible distortion. Subband coding applications to magnetic recording and transmission are discussed in some detail. Performance bounds for this type of subband coding systems are derived.**

I. Introduction

AUDIO source coding has been receiving attention in a number of ways. Many papers, e.g., [1]–[5] have been published on this topic. Presently, two standardization activities for audio bit rate reduction are taking place. One is conducted by the International Standardization Organization (ISO) [6]. The other is the European project EUREKA 147, aiming to define a broadcasting standard for digital audio. Some of the source coding methods described in the papers mentioned above have been submitted as standard proposals to ISO or EUREKA 147. Finally, Philips has recently announced a digital compact cassette (DCC) player capable of recording source coded high-quality digital audio signals on a compact cassette [7].

The audio source coding methods described in the various papers can be divided into the following categories: subband coding, e.g., [1]–[3]; and transform coding, e.g., [4], [5]. Although these papers describe different coding principles, they have three things in common. First, all papers discuss so-called waveform coders [8]. This means that in some sense they try to approximate the original input waveform and that the coding error is the consequence of additive quantization noise. Second, they all claim to achieve (near) compact disc quality at bit rates down to 96 kb/s. For an input signal taken from compact disc, with a bit rate of 705.6 kb/s, this corresponds to a reduction factor of 7.35 achieved without loss of quality. Third, all coding methods described rely heavily on human perception, more precisely on simultaneous masking. This is the psychoacoustic phenomenon that a weak signal, e.g., quantization noise, is made inaudible (masked) by a stronger signal, e.g., a pure tone. The masking signal is called the masker; the masked signal is called the target. In order to be masked, the target's level must be below the so-called masking threshold.

All papers mentioned above exploit masking in a roughly similar way. Based on an estimate of the signal's short-time power spectral density (PSI) [9], an estimate is made of the masking threshold as a function of frequency. The signal is subsequently transformed, by either subband filters or an orthogonal transform, quantized, and coded. Quantization and coding of subband signals or transform coefficients reduce the bit rate, but also introduce coding errors. The coding systems try to keep these coding errors below the masking threshold at all frequencies in the audio band. Computation of the masking threshold is typically done every 10–30 ms in order to track changes in the signal's short-time PSD.

The approach of the above papers to try to keep coding errors below the masking threshold at all frequencies seems to be based on an incorrect interpretation of the masking threshold. The masking thresholds used are mainly derived from the Zwicker–Feldtkeller curves presented in [10]. Those curves are only valid for one single tonal target. Masking may not occur when several, possibly nontonal, targets are presented simultaneously at levels at which they are individually masked, which happens when subband samples or transform coefficients are quantized. This is recognized in [4], where it is suggested that the masking threshold should be deconvolved. It is argued there that deconvolution is ill-conditioned and a suboptimal renormalization is proposed. The reason that in other papers excellent results are also obtained is probably that the masking thresholds used are chosen on the conservative side. This paper presents an alternative for the deconvolution proposed in [4] that is not suboptimal nor ill-conditioned.

It is investigated under which conditions well-defined multiple noise targets due to quantization are masked by a given audio signal. This leads to a constraint on the target levels, called the masking constraint. The masking constraint leaves freedom to choose target levels. Since the objective of audio source coding is to achieve a low bit rate, it seems sensible to choose them in such a way that they also minimize bit rate. Under the assumption that the audio signal is Gaussian and stationary, the relation between given target levels and bit rate can be established via the rate distortion theory. Finding target levels that minimize bit rate under the masking constraint is called the audio coding optimization problem.

The goal of this paper is to introduce and solve the au-

Manuscript received September 1990; revised July 21, 1991.

The author is with Philips Research Laboratories, Eindhoven, The Netherlands.

IEEE Log Number 9104238.

dio coding optimization problem. The resulting target levels are used to give estimates of lower bounds to the bit rates for audio signals. Next to that, they can also be used in audio source coding systems. This paper is a first attempt to formulate and solve the audio coding optimization problem. Therefore, results are not definite. They are based on simple psychoacoustic assumptions and derived for stationary Gaussian signals. More elaborate models can be and should be incorporated to improve estimates of bit rates. Nevertheless, this is a promising approach, since it inherently takes into account that multiple targets must be masked and thus in the end will ensure a higher coding quality and reliable bit rate estimates. The following paragraphs present a brief overview of this paper.

Since masking plays an important role in audio coding systems, it is reviewed in Section II. In the same section it is pointed out more clearly why masking thresholds as such cannot be used directly. The problem of masking multiple targets is approached by another psychoacoustical concept: the excitation pattern model, which can be seen as the underlying model of masking. It is explained in Section III. This results in the formulation of the masking constraint.

Section IV discusses waveform coding of digital audio signals. It is assumed that the audio signal can be modeled as Gaussian and stationary and that the signal and the coding error can be characterized by their PSD's. The PSD of the coding error is called the error PSD. An expression for a lower bound to the bit rate in terms of the signal's PSD and the error PSD is presented. In Section V this expression is reformulated in terms of noise target levels. This leads to the audio coding optimization problem: target levels must minimize bit rate under the masking constraint.

The computed masked error PSD or the masked target levels can be used in either subband or transform coders to allocate bits to quantizers. Section VI focuses on subband coding. Two special versions, subband coding for DCC and for the MUSICAM [1] proposal, are discussed in more detail. Their basic coding scheme is given and estimates for bit rates achievable with this type of coding scheme are presented. Although a subjective performance evaluation of the two proposals is beyond the scope of this paper, some remarks on their performance are given. Section VII contains a discussion. It summarizes the paper and tries to indicate to what extent model assumptions are valid and which points need further attention.

II. Masking

Simultaneous masking, further referred to as masking, is the phenomenon that a weak signal is made inaudible by a simultaneously occuring stronger signal. Masking is discussed in great detail in [10]–[12]. In the next two paragraphs results from [11] are repeated to explain masking.

Consider a pure tone as the target. It is inaudible if its sound pressure level (SPL) [12] is below a threshold of hearing, called the threshold in quiet. This threshold is a function of frequency. It is the bottom curve shown in Fig. 1.[1] In the presence of a second, stronger signal, the threshold of hearing differs from the threshold in quiet. It is raised for frequencies close to the frequency of the stronger signal. The new threshold is the masking threshold. Targets with levels below it are masked. The masking threshold depends on the sound pressure level and the frequency of the masker. Fig. 1 shows masking thresholds of narrow-band noise maskers with a bandwidth of 90 Hz, centered at 1 kHz, at sound pressure levels L_N = 100 dB and L_N = 60 dB. Fig. 2 depicts masking thresholds of 1 kHz pure tone maskers at sound pressure levels L_S = 90 dB and L_S = 70 dB as a masker. Fig. 3 shows masking thresholds of narrow-band noise maskers centered at frequencies f_m = 250 Hz, f_m = 1 kHz, and f_m = 4 kHz.

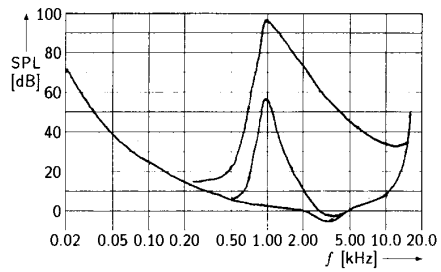


Fig. 1. Threshold in quiet (lower curve) and masking thresholds of narrow-band noise maskers centered at 1 kHz at sound pressure levels L_N = 100 dB (upper curve) and L_W = 60 dB (middle curve) for tonal targets.

Masking of a pure tone by another pure tone or by a narrow-band noise signal has been reviewed briefly to illustrate the masking effect. The masking thresholds of Figs. 1–3 may not be directly suited for use in coding systems for three reasons discussed in the following paragraphs.

The first reason for the unsuitability of the masking thresholds is that they describe masking of tonal targets. In a waveform coder the targets will mainly be noise targets, due to quantization. Very little relevant information on masking of noise targets seems present in the literature, e.g., [13]. First a definition of the noise target is required. As the ear seems to integrate over limited frequency regions called critical bands [12], it seems sensible to define critical bands of noise as targets. The definition of critical bands given in [12] is used here. According to this definition, critical bands below 500 Hz are 100 Hz wide, and above 500 Hz the critical bandwidth is approximately a third octave. Consecutive critical bands are numbered from 1 to N. For audio signals sampled at a rate f_s = 44.1 kHz, 26 critical bands have to be taken into account. The critical band scale is also used as a measure of frequency, called critical band rate z. The corresponding unit is the bark. Table I lists cutoff frequencies of critical bands.

[1]Figs. 1, 2, and 3 are derived from figures in [10].

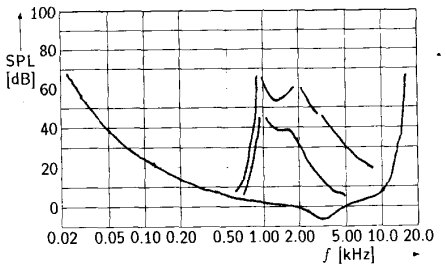


Fig. 2. Masking thresholds of pure tones of 1 kHz at sound pressure levels L_S = 90 dB (upper curve) and L_S = 70 dB (middle curve) for tonal targets.

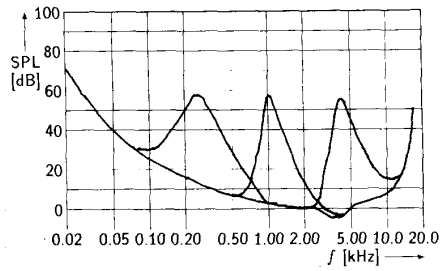


Fig. 3. Masking thresholds of narrow-band noise signals centered at frequencies f_m = 250 Hz (left-hand curve), f_m = 1 kHz (middle curve), f_m = 4 kHz (right-hand curve) for tonal targets.

TABLE I
CRITICAL BAND FREQUENCIES, TAKEN FROM [14][1]

Rate [Barks]	Δf [Hz]	f_l [Hz]	f_u [Hz]
1	—	—	100
2	100	100	200
3	100	200	300
4	100	300	400
5	110	400	510
6	120	510	630
7	140	630	770
8	150	770	920
9	160	920	1080
10	190	1080	1270
11	210	1270	1480
12	240	1480	1720
13	280	1720	2000
14	320	2000	2320
15	380	2320	2700
16	450	2700	3150
17	550	3150	3700
18	700	3700	4400
19	900	4400	5300
20	1100	5300	6400
21	1300	6400	7700
22	1800	7700	9500
23	2500	9500	12000
24	3500	12000	15500
25	6550	15500	19500
26	6550	19500	24600

[1]Reprinted with permission from Academic Press.

A critical band noise target has a flat PSD within one critical band and a zero PSD outside. In fact, because of the integration one would expect the masking thresholds for critical band noise targets to be similar to those for tonal targets. Fig. 4 shows a masking threshold of a tonal masker of 400 Hz for critical band noise targets as a function of the critical band number. It is the result of an informal experiment, with the author as subject. It confirms the expectation that masking thresholds for critical band noise targets are similar to those for tonal targets. Although further study of masking thresholds for critical band noise targets and tonal and noise maskers is needed, it is further assumed that the thresholds for tonal targets can be used instead. A second conclusion from the fact that the masking curves for tonal and critical band noise targets do not differ very much, is that the exact spectral shape of the noise target within the critical band is not important. This is convenient because the noise targets in this paper will be due to quantization noise, which cannot always be considered spectrally flat.

The second reason for the unsuitability of the masking thresholds is that they describe masking by only one masker. Audio signals consist of many maskers. What is required is an addition law for masking thresholds. Results on addition of masking thresholds are given in, e.g., [14], [15]. Two principles are commonly used in coding to obtain a total masking threshold from individual ones. The most conservative one is to define the total masking threshold at a certain frequency as the maximum of all individual masking thresholds and the threshold in quiet

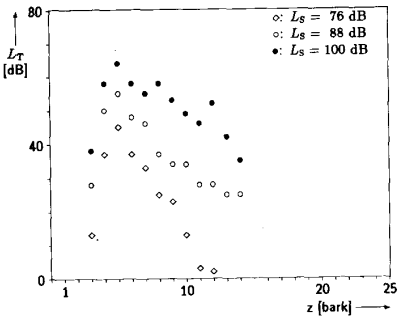


Fig. 4. Masking thresholds L_T of critical band noise targets with a 400 Hz masker at various sound pressure levels L_S.

at that frequency [3]. The other approach is to assume that masking is additive, so that the total masking threshold is obtained as the sum of the individual thresholds and the threshold in quiet [2], [4]. Before the individual masking thresholds can be computed, the maskers in the signal have to be identified. Usually this is done by analyzing an estimate of the signal's short-time PSD. As a crude estimate, signal powers in subbands can be used, as is done in [2]. More elaborate estimates based on a discrete Fourier transform are also used [1]. In a transform coder the estimate can be derived from the transform coefficients [4].

The third reason for the unsuitability of the masking thresholds is that they describe masking for only one target. The result of a simple experiment suffices to dem-

onstrate that these masking thresholds are not correct for multiple targets, although it is often assumed otherwise [1]–[3]. In the experiment the masker was presented together with all targets in critical bands 3 to 12. The targets were presented at levels at which they are individually masked. It was found in that case that the noise was not masked. The levels of the targets had to be reduced by 9 dB to reobtain masking.

The first two problems have been overcome by, respectively, defining the proper targets and by assuming an addition law for masking thresholds. The problem of masking multiple critical band noise targets is further tackled in Section III.

III. Masking Multiple Targets

To deal with the problem of masking multiple targets, the concept of the excitation pattern is used. The excitation pattern model [16], [17], [12], [18] explains discrimination between sounds by the ear and can be seen as the basis of masking.

According to [12], discrimination is based upon changes in patterns of neural activity. Neurons are organized in such a way that groups of neurons correspond to frequency regions. Excitation with a pure tone in one critical band will activate neurons corresponding to that particular critical band, but it will also activate neurons corresponding to frequency regions outside of it. The response to the excitation as a function of critical band rate can be described by an excitation pattern. This has a shape similar to that of a masking threshold [16]. The excitation pattern is usually presented as an excitation level on a decibel scale and as a function of critical band rate. It is maximal in the critical band where the excitation takes place. The excitation level in that critical band is, by definition [16], equal to the sound pressure level of the exciting signal.

Maskers and targets together determine the excitation pattern. According to [16], detection of a target occurs if in any critical band the excitation pattern of masker and target together differs more than 1 dB from the excitation pattern of masker alone. According to [17], detection of a target occurs if these differences summed over all critical bands are greater than a certain threshold. In this paper the approach of [16] is followed, but the results can be adapted to the approach of [17]. The excitation pattern model explains the outcome of the experiment of the last paragraph of Section II: the excitation pattern of a single target and masker differs less than 1 dB from the masker's excitation pattern, whereas the excitation pattern of multiple targets and masker differs more. This also gives a clue as to how to compute the target levels in order to obtain complete masking. They must be chosen such that they do not change the excitation pattern by more than 1 dB in any critical band. The idea of this section is to develop an expression for the change in the excitation pattern due to critical band noise targets. Then this change is limited to 1 dB resulting in the masking constraint on the target levels.

The following expression gives a mathematical model for the excitation pattern used throughout this paper

$$e(\sigma^2[j], i, j) = \begin{cases} \sigma^2[j]\beta^{|j-i|}, & i < j, \\ \sigma^2[j], & i = j, \\ \sigma^2[j]\alpha^{|j-i|}, & i > j. \end{cases} \quad (1)$$

In this expression $e(\sigma^2[j], i, j)$ is the excitation in critical band i, $1 \le i \le N$, caused by a signal with power $\sigma^2[j]$ in critical band j, $1 \le j \le N$. The α and β define the slopes of the excitation pattern. The slope towards higher critical bands usually depends on the excitation level. For excitation levels in the area of 80 dB, $\alpha = 0.15$, corresponding to a slope of 8 dB/bark and for excitation levels around 50 dB, $\alpha = 0.25$, corresponding to 6 dB/bark. The slope towards lower critical bands is constant and well-approximated by $\beta = 0.003$, corresponding to 25 dB/bark. Plotted on a decibel scale excitation patterns according to (1) resemble the masking thresholds of Fig. 1. The difference is an upward shift in level.

The frequency resolution of the excitation pattern in (1) is restricted to one critical band. Refinements, based on a more precise spectral analysis of the signal, would result in a more accurate computation of the excitation pattern. However, this would lead to a more complicated analysis, which is beyond the scope of this paper.

In order to be able to compare excitation patterns of multiple maskers and targets, a rule of addition is needed to compute a total excitation pattern from individual ones. Because of the relation between masking and excitation patterns, it is assumed that the rules of addition for masking thresholds also apply for excitation patterns. The rule of addition used here is the modified power law discussed in [15]. According to this rule, the sum e_S of two excitations in a critical band e_A and e_B is given by

$$e_S = (e_A^p + e_B^p)^{1/p} \quad (2)$$

where p, $0 < p < 1$, is the power law constant. The choice $p = 1$ results in the pure addition of excitation patterns. In experiments described in [15] p is fitted to data resulting in values in the range 0.1–0.5.

The excitation in critical band i is given by

$$e[i] = ((e_m[i])^p + (e_t[i])^p)^{1/p} \quad (3)$$

where $e_m[i]$ is the total excitation pattern in critical band i of the maskers and $e_t[i]$ is the total excitation pattern in critical band i of the critical band noise targets. According to the addition rule in [15], the excitation of the maskers is given by

$$e_m[i] = \left((e_Q[i])^p + \sum_{j=1}^{N} (e(\sigma_m^2[j], i, j))^p \right)^{1/p} \quad (4)$$

where $\sigma_m^2[j]$ is the signal power in critical band j and $e_Q[i]$ is the excitation in quiet. The excitation in quiet is simply a level-shifted version of the threshold in quiet. The ex-

citation of the targets is given by

$$e_t[i] = \left(\sum_{j=1}^{N} (e(\sigma_t^2[j], i, j))^p \right)^{1/p} \quad (5)$$

where $\sigma_t^2[j]$ is the, yet unknown, target level in critical band j.

According to the detection rules of [16], masking occurs if

$$10 \log \left(\frac{((e_t[i])^p + (e_m[i]^p)^{1/p}}{e_m[i]} \right) < 1, \quad i = 1, \cdots, N \quad (6)$$

or, equivalently,

$$(e_t[i])^p < (10^{p/10} - 1)(e_m[i])^p, \quad i = 1, \cdots, N. \quad (7)$$

The choice $p \doteq 0.48$ leads to a ratio $e_t[i]/e_m[i]$ of -20 dB, roughly corresponding to the masking at the top of a masking threshold as depicted in Fig. 1.

The signal powers within critical bands $\sigma_m^2[j]$ can be derived from an estimate of the signal's short-time PSD. Therefore, the right-hand side of (7) can be computed by substituting the $\sigma_m^2[j]$ into (4). Let $\boldsymbol{m}$ denote a vector derived from the right-hand side of (7), with elements

$$m_i = (10^{p/10} - 1)(e_m[i])^p, \quad i = 1, \cdots, N. \quad (8)$$

It is more complicated to set up a manageable left-hand side involving the $\sigma_t^2[i]$. First it is assumed that the excitation patterns of the targets have shapes independent of their levels. This can be justified since all targets will be relatively weak signals due to quantization noise. Furthermore, let $\boldsymbol{t}$ denote a vector derived from the powers of critical band targets $\sigma_t^2[j]$ with elements

$$t_j = (\sigma_t^2[j])^p, \quad j = 1, \cdots, N \quad (9)$$

and define elements of the $N \times N$ matrix $\boldsymbol{E}$ by

$$e_{i,j} = \begin{cases} \beta^{|j-i|p}, & i < j, \\ 1, & i = j, \\ \alpha^{|j-i|p}, & i > j \end{cases} \quad (10)$$

where α and β are the same constants as in (1). The left-hand side of (7) can now be written in a matrix form

$$\begin{pmatrix} (e_t[1])^p \\ \vdots \\ (e_t[N])^p \end{pmatrix} = \boldsymbol{Et} \quad (11)$$

and (7) can be reformulated as

$$\boldsymbol{Et} \leq \boldsymbol{m}. \quad (12)$$

The $\leq$ sign must be interpreted componentwise. Equation (12) is the masking constraint on the elements of $\boldsymbol{t}$, or via (9) on the $\sigma_t^2[j]$. All multiple critical band noise targets with levels satisfying it are masked. This still leaves freedom to choose target levels in many ways and thus, in coding systems, to allocate bits to quantizers in many ways. Therefore, in Section IV an additional requirement on target levels to minimize the bit rate is formulated. Together with the masking constraint (12) this defines the audio coding optimization problem.

In [4] it is mentioned that the masking threshold must be deconvolved in order to obtain masked noise levels and that this is an ill-conditioned process, leading to possibly negative power values. Therefore, a suboptimal solution is proposed. In terms of this paper, this deconvolution problem is the problem of solving (12) with equality. Indeed, the matrix $\boldsymbol{E}$ may be ill-conditioned or $\boldsymbol{m}$ may not be an image of a vector with only positive elements, in which case strange results can be expected. In this paper these problems are evaded by regarding (12) as constraints to an optimization problem.

IV. Waveform Coding of Audio Signals

The sampling rate f_s of a digital audio signal should be high enough to ensure an audio bandwidth wide enough for HiFi quality. In this paper it is 44.1 kHz. Although in practice the audio samples are 16-bit integers, they are assumed to be a real number for convenience of analysis.

Let the digital audio signal be denoted by $s[i]$, $i = -\infty, \cdots, +\infty$. It is assumed to be a random process that is at least wide-sense locally stationary [19], such that a meaningful estimate of a short-time PSD can be made. It is also assumed that $s[i]$ is a zero-mean Gaussian process. The output of the decoder is $\hat{s}[i]$, $i = -\infty, \cdots, +\infty$. The coding error $\epsilon[i]$ is defined by

$$\epsilon[i] = \hat{s}[i] - s[i], \quad i = -\infty, \cdots, +\infty. \quad (13)$$

The coding error is characterized by its PSD, called the error PSD $S_{\epsilon\epsilon}(\exp(j\theta))$, $-\pi \leq \theta \leq \pi$. The signal's PSD is denoted by $S_{ss}(\exp(j\theta))$, $-\pi \leq \theta \leq \pi$. The definition of PSD used in this paper is the one for discrete-time signals. It gives the sample power per unit of normalized frequency. The continuous-time PSD, giving the power per unit of frequency after *D/A* conversion, can easily be derived from it. For the error PSD, for instance, it is given by

$$S_{\epsilon\epsilon}^c(j2\pi f) = \begin{cases} \frac{1}{f_s} S_{\epsilon\epsilon}\left(\exp\left(j2\pi \frac{f}{f_s}\right)\right), \\ \qquad -f_s/2 \leq f \leq f_s/2, \\ 0, \quad \text{otherwise.} \end{cases} \quad (14)$$

The bit rate R needed to code a signal with PSD $S_{ss}(\exp(j\theta))$, resulting in an error PSD $S_{\epsilon\epsilon}(\exp(j\theta))$, satisfies

$$R \geq \frac{1}{\pi} \int_0^{\pi} \max\left(0, \frac{1}{2} \log_2 \left(\frac{S_{ss}(\exp(j\theta))}{S_{\epsilon\epsilon}(\exp(j\theta))} \right)\right) d\theta \text{ b/sample}. \quad (15)$$

This bound can be derived using results presented in [20]. Two observations can be made from (15). First, from the

max operator it follows that frequency regions where the error PSD is greater than the signal's PSD do not contribute to the rate. In a good coding system, data describing the signal in these frequency regions are not transmitted. Second, the bit rate is determined by the desired signal-to-noise ratio as a function of frequency. The bound (15) illustrates that it is of importance to have an accurate estimate of the masked error PSD because it directly influences the bit rate.

The integral (15) is a bound that may be difficult to reach. A coding system using scalar quantizers, as is usually the case in audio source coding, cannot perform better than about 0.25 b above the bound (15) [21], [22]. A coder that tries to approach this bound more closely than that has to make use of vector quantizers [22]. Moreover, (15) is derived for stationary signals. This is an unrealistic assumption. In a practical application the encoder and decoder will try to adapt to changes in the signal's PSD. This increases the bit rate, because the encoder must inform the decoder about these changes by transmitting side information. For instance, in the MUSICAM system the side information consists of scale factors and bit-allocation information for the block-companding quantizers [8], and requires up to 0.20 b/sample.

The above paragraph explains that bit rates achieved in practice may be higher than (15). However, there is also an argument from which it follows that sometimes they may be lower. This is because (15) is based on the assumption that the audio signal is Gaussian. This assumption is not always true, particularly not when the music signal consists of pure tones. In these cases a substantially lower bit rate can be obtained. In fact, it is shown in [21] that with respect to the mean-squared error criterion, non-Gaussian stochastic processes have a rate distortion function that lies below that of a Gaussian process. This implies that for non-Gaussian music signals and a given error PSD $S_{\epsilon\epsilon}(\exp(j\theta))$, the achievable minimum bit rate is less than predicted by (15).

In spite of the fact that for non-Gaussian signals the minimum achievable bit rates may be lower than (15), this expression will be used in the remainder of this paper. The reasons are that even though it may not be completely correct, at least it gives some indication about bit rates required. Also, it may be assumed that at least sometimes music can be modeled correctly as Gaussian.

Figs. 5 and 6 show estimates of lower bounds to the bit rate as a function of time for two fragments of about 10 s of tubular bells and a symphony orchestra, respectively. The error PSD was fixed as $S_{\epsilon\epsilon}(\exp(j\theta)) = 1/12$, corresponding to the white quantization noise present in compact disc signals. The reason to give these examples is to demonstrate that coding at the desired bit rates of 2–4 b/sample is impossible if the required CD quality is defined as a mean-squared error equal to the quantization error of compact disc signals. The signal's short-time PSD was repeatedly estimated by a 1152 point discrete Fourier transform. The integral (15) was approximated by a discrete sum over the 577 frequency samples at which the

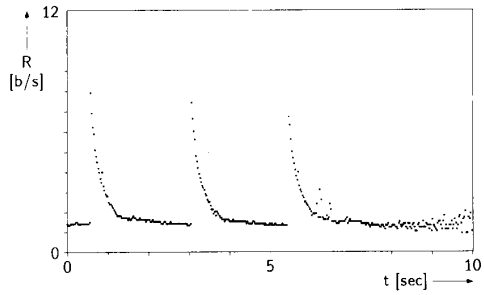


Fig. 5. Lower bound to the bit rate in bits per sample for a fragment of 10 s of tubular bells. Coded with a flat error PSD $S_{\epsilon\epsilon}(\exp(j\theta)) = 1/12$.

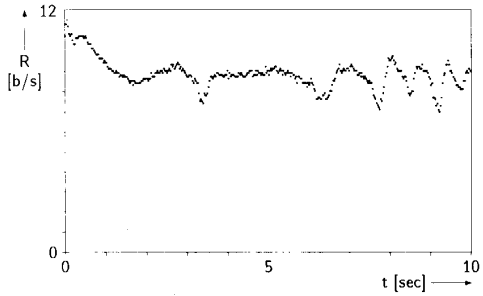


Fig. 6. Lower bound to the bit rate in bits per sample for a fragment of 10 s of symphony orchestra. Coded with a flat error PSD $S_{\epsilon\epsilon}(\exp(j\theta)) = 1/12$.

short-time PSD was estimated. The curves give lower bounds for bit rates, expressed in bits per sample, of codecs coding at compact disc quality without exploiting masking. The bit rate is mostly above the desired range of 2–4 b/sample.

The bit rate required for the fragment of tubular bells is generally low. Only during and shortly after a beat on the instrument is it substantially higher. This occurs three times in Fig. 5. The reason for the increased bit rate during and after a beat is that then the signal has many spectral components and a large part of the signal's PSD will be greater than the error PSD. Most of these components die out after the beat. The symphony orchestra has a filled PSD for most of the time and requires a much more constant bit rate.

Graphs of estimates of bounds such as Figs. 5 and 6, and those presented later must be interpreted carefully, even if the signal is Gaussian. There are actually two reasons why in that case the bounds cannot easily be reached. The first one is the necessity of transmitting side formation as has already been mentioned. The second is that the bounds have been computed by cutting the signal into blocks. Each block is assumed to be a segment of a realization of a stationary stochastic Gaussian process. From this block an estimate of the bit rate for the entire process is made. To actually reach that bit rate, it may be necessary to quantize and code blocks larger than the one used to make the bit-rate estimate.

In the following section (15) will be expressed in terms

of target levels and combined with the masking constraint (12), thus leading to the audio coding optimization problem.

V. Audio Coding Optimization Problem

The target PSD of critical band noise targets $T(\exp(j\theta))$ is given by

$$T(\exp(j\theta)) = \frac{f_s}{2\Delta f_i}\sigma_i^2[i] = \frac{f_s}{2\Delta f_i}t_i^{1/p},$$

$$\frac{2\pi f_{l,i}}{f_s} \le \theta < \frac{2\pi f_{u,i}}{f_s} \qquad (16)$$

where t_i is defined in (9), Δf_i, $f_{l,i}$, and $f_{u,i}$ are the bandwidth and the lower and upper cutoff frequencies of critical band i, respectively; see Table I. Define the fraction w_i of signal bandwidth, $f_s/2$ occupied by critical band i by

$$w_i = \frac{2\Delta f_i}{f_s}. \qquad (17)$$

On substitution of (16) and (17) into (15) and after ignoring the max operator, it follows that optimal t_i have to be chosen such that

$$-\frac{1}{\pi}\int_0^\pi \frac{1}{2}\log_2\left(T(\exp(j\theta))\right)\,d\theta = -\frac{1}{2}\sum_{i=1}^{N} w_i \log_2\left(\frac{t_i^{1/p}}{w_i}\right) \qquad (18)$$

is minimized under the constraints (12). Instead of minimizing (18) it also suffices to minimize the objective function

$$Q(t) = -\sum_{i=1}^{N} w_i \ln(t_i) \qquad (19)$$

under the constraints (12)

$$\boldsymbol{Et} \le \boldsymbol{m}.$$

This is the mathematical formulation of the audio coding optimization problem. Solving it yields target levels that are masked and at the same time minimize bit rate. It is a nonlinear optimization problem that can be solved with commercially available mathematical libraries. The optimal $\sigma_i^2[i]$ and the optimal error PSD can then be computed according to (16). The resolution of the error PSD is limited to critical bands. It can be further improved by defining narrower noise targets. However, since both the computation of the total masking threshold as well as the computation of the target levels already have a strong smoothing effect, it is questionable whether this improvement would be really substantial. Also, since what is obtained is a stepwise approximation of the error PSD, it can be smoothed by a suitable operator. The above procedure for estimating masked error PSD's or masked target levels can be used in two ways: to derive performance bounds for audio coding or to compute masked noise levels in coding systems. Both are briefly discussed below.

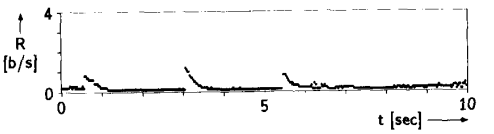


Fig. 7. Lower bound to the bit rate in bits per sample for a fragment of 10 s of tubular bells. Coded with full exploitation of masking.

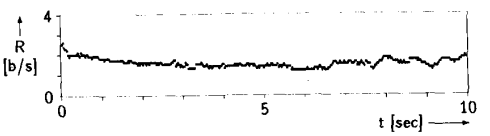


Fig. 8. Lower bound to the bit rate in bits per sample for a fragment of 10 s of symphony orchestra. Coded with full exploitation of masking.

The procedure to estimate performance bounds is similar to the one followed to derive Figs. 5 and 6, only now, instead of choosing $S_{\epsilon\epsilon}(\exp(j\theta)) = 1/12$, the optimal masked error PSD is derived by solving the audio coding optimization problem. Figs. 7 and 8 show estimates of lower bounds to the bit rates obtained by minimizing (19) under constraint (12) for the same two fragments as used before. The excitation pattern model used is a simple version of (1) with $\alpha = 0.25$, $\beta = 0.003$, and $p = 0.48$. The results show that the use of masking is fruitful since both estimates are substantially lower than the ones shown in Figs. 5 and 6. In addition to remarks on bounds previously made, it must be said that the results are as good as the excitation pattern model used to derive them, which is just a simple one. More accurate models, including more accurate computation of excitation patterns and better addition rules will yield more accurate bounds. Nevertheless, the results clearly show the improvement obtained by exploiting masking and they at least suggest that more can be achieved than the not really transparent quality at 2 b/sample of present systems.

It is interesting to illustrate how target levels can depend on the signal's spectral behavior. Therefore, as a further illustration estimates of the signal levels in critical bands $\sigma_m^2[j]$ and of target levels $\sigma_t^2[j]$ obtained from 1152 samples of the fragments of tubular bells and symphony orchestra have been plotted in Figs. 9 and 10, respectively. The levels are plotted on a decibel scale. The reference is the noise level of compact disc signals, which is equal to 1/12. These figures clearly illustrate the different spectral characteristics of both fragments. Most of the power of the fragment of tubular bells is concentrated in a small number of critical bands, whereas the power of the fragment of symphony orchestra is more equally distributed over the critical bands. The target levels in the fragment of tubular bells are much closer to or even above the signal levels than is the case for the fragment of symphony orchestra. This explains that the required bit rate for the fragment of tubular bells is much smaller.

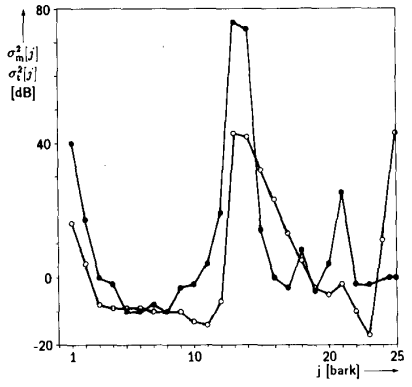


Fig. 9. Signal levels $\sigma_m^2[j]$, ●, and target levels $\sigma_t^2[j]$, ○, estimated from 1152 samples of tubular cells.

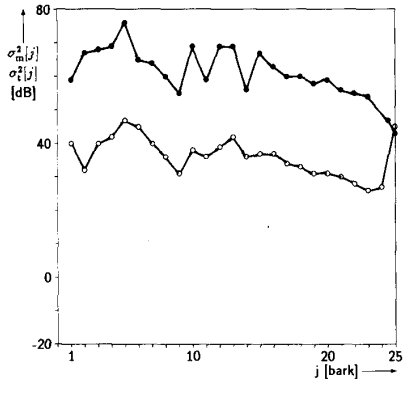


Fig. 10. Signal levels $\sigma_m^2[j]$, ●, and target levels $\sigma_t^2[j]$, ○, estimated from 1152 samples of symphony orchestra.

The above procedure for estimating masked error PSDs or masked target levels can also be used in a practical application. Before this is discussed in a little more detail, its usefulness must be established. There are three possible objections. First, remember that vector quantizers are required to approximate the bit rate predicted by (15) any closer than 0.25 b/sample. Practical audio coding systems, with the exception of low rate speech coders, do not often contain vector quantizers, but merely use scalar, uniform quantizers. Moreover, they have to adapt to changes in signal statistics. Therefore, the bit rate may come out higher than predicted by (15). Second, for non-Gaussian signals the bound may be substantially lower than that. Third, in practical situations it is seldom of interest to code at a bit rate that is varying. In the transmission and recording applications discussed here, a fixed bit rate is required.

With respect to the first two objections, it can be said that for a very large class of uniformly quantized signals, the quantization errors variance decreases exponentially with the number of bits. Essentially, this is the rule of thumb that signal-to-noise ratio decreases 6 dB/b. This implies that if scalar uniform quantizers are used and in the case of non-Gaussian signals, it still makes sense to minimize (18), although (15) does not accurately predict the achievable bit rate. With respect to the often required fixed bit rate, there are two solutions. The first is to code at a varying bit rate and use a buffer to make it fixed. This is not often done in music nor in speech coding and will not be considered here. The second solution is still to compute the optimal masked error PSD, but to use it as a spectral error-weighting function. This implies that the coding error will have a PSD with the shape of the optimal masked error PSD, but it will be scaled upward or downward, when the required bit rate is, respectively, greater or less than the fixed bit rate. This approach is followed in the adaptive bit-allocation procedure described in [2].

The practical coding procedure is outlined as follows. The first step is to obtain, e.g., via an estimate of the signal's short-time PSD, estimates of the excitation levels $e_m[i]$ in critical bands. These are used in (8) to obtain the right-hand side of the masking constraint (12). The next step is to solve the audio coding optimization problem (19). In a forthcoming paper an efficient solution to it will be presented. From the computed masked target levels a masked error PSD can be computed by (16). Either the masked error PSD or the masked target levels can be used to compute the number of bits needed to quantize subband signals or transform coefficients. An audio coding scheme computing the masked error PSD and using it in combination with an adaptive bit allocation method will be presented in a forthcoming paper.

The following section describes how masking is applied in subband coding. The example given is the basis for MUSICAM as well as DCC. Lower bounds to the bit rates achievable with these systems are given and compared to the results of this section.

VI. Application to Subband Coding

In a subband coder, a filterbank splits the audio signal into a number of adjacent frequency bands called subbands. The sampling rate of the subbands is reduced to a fraction of the input sample rate, in such a way that the sample rate of each subband is twice its bandwidth. Subband signals are quantized and coded. By means of another filterbank, the decoder merges the quantized subband signals into a reconstruction of the input signal. In this manner, after reconstruction the quantization errors remain in the subband in which they were introduced. Careful design of the filterbanks can guarantee a perfect [23] or almost perfect reconstruction of the input signal if quantization is omitted. A good overview of design methods for filterbanks is given in [24].

In subband coding proposals DCC and MUSICAM filterbanks, based on proposals in [25], [26], are used to split the audio signal into 32 equally spaced subbands. Fig. 11 gives a simplified basic scheme. The boxes LP, BP, and HP are lowpass, bandpass, and highpass filters, respectively. In the encoder they are followed by decimators. In the decoder they are preceded by interpolators

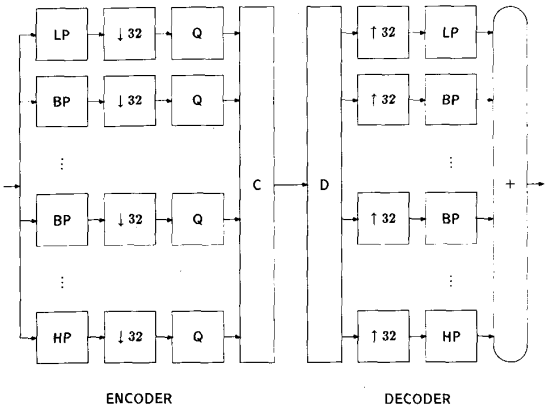


Fig. 11. Basic scheme of DCC and MUSICAM subband coder.

[24]. In the discussion that follows the filters are assumed to be ideal bandpass filters.

The boxes Q denote quantizers. Quantization is adaptive in two ways. In the first place, it is adaptive to the local signal level. Before quantization, which is uniform [8], the subband signals are divided into blocks of 12 samples which are scaled in such a way that the samples are in the range $[-1, 1]$. The scale factors are coded and transmitted as side information. This method of quantization is known as block-companding or gain-adaptive quantization [8]. In the source decoder, the quantized samples are multiplied by the scale factors in order to obtain replicas of the subband samples. In the second place quantization is adaptive since the numbers of quantization levels are chosen in such a way that an attempt is made to keep the quantization noise masked. In principle, the numbers of quantization levels are recomputed for every group of 32 blocks of 12 subband samples, called an allocation window. It corresponds to approximately 8 ms. The numbers of quantization levels are transmitted as side information.

The boxes C and D in Fig. 11 perform additional encoding and decoding of quantized samples, scale factors, and allocation information.

Differences between DCC and MUSICAM mainly concern the way the allocation information is computed and the boxes C and D. The bit rate for DCC is 192 kb/s. For MUSICAM it is either 128, 96, or 64 kb/s. Therefore, more accurate computation of masking thresholds and more effective coding of quantized samples, scale factors, and allocation information are required. Only computation of masking thresholds is discussed in the following paragraphs. First the ideal procedure of computing the masked noise power in each subband is discussed, then the practical solutions of DCC and MUSICAM are treated in some detail.

Ideally, the procedure would be as follows. Quantizers are assumed to produce an uncorrelated quantization error which has a flat PSD within the subband it is introduced in. The quantization noise power is assumed to be given by

$$\sigma_q^2 = \frac{\Delta^2}{12} \tag{20}$$

where Δ is the quantization step size. These assumptions are justified if the number of quantization levels is large enough, if Δ is sufficiently small, and if quantizer overload distortion is negligible [8]. Overload distortion does not occur, since block-companding quantizers are used. For a small number of quantization levels the error may become correlated, but (20) remains a useful approximation of the quantization error variance.

The number of subbands is further denoted by K. Starting from the masked error PSD $S_{\epsilon\epsilon}(\exp(j\theta))$, the masked noise power $\sigma_n^2[i]$ in subband i can be taken as

$$\sigma_n^2[i] = \frac{1}{K} \min_{(i-1)(\pi/K) \le \theta < i(\pi/K)} S_{\epsilon\epsilon}(\exp(j\theta)). \tag{21}$$

Once the masked noise power in each subband is known, quantizers can be adapted in such a way that the power of the quantization noise is just below it. If the number of quantization levels in subband i and the scale factor are denoted by l_i and p_i, respectively, then the quantization step size is given by

$$\Delta = \frac{2p_i}{l_i}.$$

The quantization noise power in that subband is masked if

$$\frac{1}{12}\left(\frac{2p_i}{l_i}\right)^2 \le \sigma_n^2[i] \tag{22}$$

from which the correct l_i can be computed. Unfortunately, this would lead to a varying bit rate

$$R = \sum_{i=1}^{K} \log_2 (l_i) + R_s \tag{23}$$

where R_s is the bit rate required for side information. As has already been remarked in Section V, a varying bit rate is undesirable. A solution is to use the optimal error PSD as a weighting function. In [2] this is done by minimizing the sum of the so-called noise-to-mask ratios

$$\sum_{i=1}^{K} \frac{1}{12}\left(\frac{2p_i}{l_i}\right)^2 \frac{1}{\sigma_n^2[i]} \tag{24}$$

under the constraint of a fixed bit rate.

It is interesting to know the theoretically achievable bit rate of a subband coder. Assume that subband signals are zero-mean Gaussian, with a variance $\sigma_s^2[i]$, given by

$$\sigma_s^2[i] = \frac{1}{\pi} \int_{i(\pi/K)}^{(i+1)(\pi/K)} S_{SS}(\exp(j\theta))\, d\theta. \tag{25}$$

In a subband coder, no use is made of knowledge about the short-time PSD within the subband. Therefore, to estimate a bound to the bit rate, it is assumed that each sub-

band is an uncorrelated Gaussian source. This means that the total bit rate satisfies [21]

$$R \geq \frac{1}{K} \sum_{i=1}^{K} \max\left(0, \frac{1}{2}\log_2\left(\frac{\sigma_s^2[i]}{\sigma_n^2[i]}\right)\right) \text{ b/sample.} \quad (26)$$

Figs. 12 and 13 show estimates for the bounds (26) for the two fragments of Section IV. The bounds are about 0.5 b/sample higher than the bounds presented in Figs. 7 and 8, and are often around 2 b/sample. Considering again that realistically achievable bit rates may be somewhat above this bound, it becomes questionable whether transparent quality can be achieved with a such a subband coder at bit rates below 96 kb/s. Even transparent quality at 96 kb/s could be difficult to achieve. If more subbands are used, (26) starts to approximate the integral (15), so it can be expected that in that case the bound will decrease. This also illustrates that the loss of 0.5 b/sample is due to the loss of spectral resolution, reflected in the assumption that each subband is considered as an independent Gaussian source. If the spectral structure within a subband can be better exploited, e.g., by adding ADPCM coding [8], the bound will decrease further.

After this general discussion on subband coding and the bit rates that it can achieve, the computation of masked noise powers $\sigma_n^2[i]$ in DCC recorders and MUSICAM is discussed in greater detail.

In a MUSICAM coder a total masking threshold is computed from an estimate of the signal's short-time PSD, computed by means of a discrete Fourier transform on the input samples. This computation is repeated every 24 ms. The optimization step in which an optimal masked error PSD is computed under the masking constraint is omitted. Instead the masked noise powers are taken as the minimum value of the masking threshold within the subband. In DCC recorders a similar approach to compute the masking threshold is followed, but because of the higher bit rate (192 kb/s) it suffices to estimate the short-time PSD directly from the subband signals.

In both cases, after computation of the masked noise powers $\sigma_n^2[i]$, bits are allocated to the quantizers by the adaptive bit-allocation algorithm of [2], that minimizes the sum of the noise-to-mask ratios (24).

A full subjective performance evaluation of DCC and MUSICAM is beyond the scope of this paper, as its main topic is the audio coding optimization problem rather than DCC or MUSICAM. A few remarks, however, can be made.

DCC has been extensively evaluated within Philips. The result of these tests was that for most fragments the subjects could not distinguish between coded and original. For some fragments some subjects could hear differences. The quality of those fragments was still judged as near CD.

MUSICAM has been evaluated with other codes in two extensive ISO listening tests at bit rates 128, 96, and 64 kb/s (sampling rate f_s = 48 kHz) and was always ranked in first or second place. At 128 kb/s original and coded

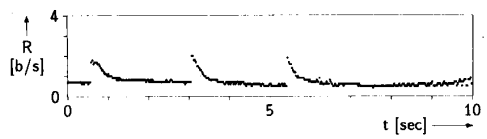


Fig. 12. Lower bound to the bit rate in bits per sample for a fragment of 10 s of tubular bells in a 32 band subband coder.

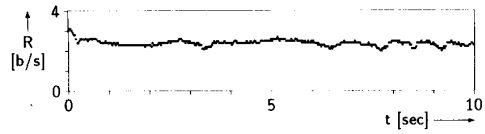


Fig. 13. Lower bound to the bit rate in bits per sample for a fragment of 10 s of symphony orchestra in a 32 band subband coder.

fragments were hardly or not distinguishable. Slight degradations were sometimes audible at 96 kb/s. This is in correspondence to the results of Fig. 13, which show that a bit rate greater than 2 b/sample sometimes is required. The quality at 64 kb/s was not always good, but according to the results presented here this could not be expected.

VII. Discussion

Psychoacoustic results such as masking and excitation pattern models have been combined with results from rate distortion theory to formulate the audio coding optimization problem. The solution of the audio optimization problem is a masked error spectrum, prescribing how quantization noise must be distributed over the audio spectrum in order to obtain a minimal bit rate and an inaudible coding error. This result cannot only be used to estimate performance bounds, but can also be directly applied in audio coding systems. Examples of subband coding systems and performance bounds for their bit rates have been given.

With respect to the psychoacoustic results that are used, it must be remarked that some assumptions had to be made about additivity of excitation patterns. The assumptions seem reasonable, but need further verification. It has also been remarked that the bounds produced are as good as the excitation pattern models used to derive them. This calls for more research on suitable psychoacoustic models for coding. It has been indicated at some points that more accurate excitation pattern models can be incorporated.

With respect to the use of the rate distortion theory, it must be noted that all results are derived for Gaussian signals and that the bounds shown are only valid if signals are Gaussian. However, this does not make the derivation useless because at least sometimes the Gaussian assumption will be realistic. In Section V it has been argued that the Gaussian assumption is not very restrictive for the use of the results in coding schemes.

There is a limitation to the results presented here. The excitation pattern models used are for stationary targets

and maskers. Transient behavior, masking shortly before, during, and after transients in audio signals, has not been discussed. It would be very interesting, but also complicated, to extend the results into this direction, especially since some of the defects heard when signals are coded at bit rates below 96 kb/s seem to be caused by poor coding at transients.

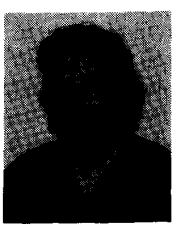
From the estimates of performance bounds shown, it can be concluded that coding at CD quality without exploiting masking is impossible, but that the use of masking is so beneficial that bit rates in the area of 2 b/sample seem feasible.

ACKNOWLEDGMENT

The author wishes to thank his colleagues R. van der Waal from Philips Research Laboratories and A. Houtsma and A. Kohlrausch from IPO who have contributed to this paper during many fruitful discussions.

REFERENCES

[1] G. Stoll and Y. F. Dehery, "High quality audio bit rate reduction system family for different applications," presented at Proc. Supercomm '90, Atlanta, GA, 1990.
[2] R. N. J. Veldhuis, M. Breeuwer, and R. G. van der Waal, "Subband coding of digital audio signals," *Philips J. Res.*, vol. 44, no. 2-3, pp. 329-343, 1989.
[3] G. Theile, M. Link, and G. Stoll, "Low bit rate coding of high-quality audio signals," *AES Preprint*, p. 2431, 1987.
[4] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 314-323, 1988.
[5] K. Brandenburg, "High quality sound coding at 2.5 bit/sample," *AES Preprint*, p. 2582, 1988.
[6] H. G. Musmann, "The ISO coding standard," presented at Proc. GLOBECOM '90, 1990.
[7] *DCC, the Fundamentals.* Philips Press Release, Philips, 1990.
[8] N. S. Jayant and P. Noll, *Digital Coding of Waveforms.* Englewood Cliffs, NJ: Prentice-Hall, 1984.
[9] S. L. Marple, Jr., *Digital Spectral Analysis with Applications.* Englewood Cliffs, NJ: Prentice-Hall, 1987.
[10] E. Zwicker and R. Feldtkeller, *Das Ohr als Nachrichtenempfänger.* Stuttgart: S. Hirzel Verlag, 1967.
[11] E. Zwicker and H. Fastl, *Psychoacoustics.* Berlin: Springer-Verlag, 1990.
[12] B. Scharf and S. Buus, "Stimulus, physiology, thresholds," in *Handbook of Perception and Human Performance, Volume 1.* New York: Wiley, 1986, pp. 14.1-14.71.
[13] R. P. Hellman, "Asymmetry of masking between tone and noise," *Perception Psychophys.*, vol. 11, no. 3, pp. 241-246, 1981.
[14] E. Zwicker and S. Herla, "Über die Addition von Verdeckungseffekten," *Acustica*, vol. 34, pp. 89-97, 1975.
[15] L. E. Humes and W. Jesteadt, "Models of the additivity of masking," *J. Acoust. Soc. Amer.*, vol. 85, no. 3, pp. 1285-1294, 1989.
[16] E. Zwicker, "Masking and psychological excitation as consequences of the ear's frequency analysis," in *Frequency Analysis and Periodicity Detection in Hearing.* Leiden: Sijthoff, 1970, pp. 376-396.
[17] M. Florentine and S. Buus, "An excitation-pattern model for intensity discrimination," *J. Acoust. Soc. Amer.*, vol. 70, no. 6, pp. 1646-1654, 1981.
[18] R. D. Patterson and B. C. J. Moore, "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selectivity in Hearing.* London: Academic, 1986, pp. 123-177.
[19] A. Papoulis, *Probability, Random Variables, and Stochastic Processes.* Tokyo: McGraw-Hill, 1965.
[20] D. J. Sakrison, "The rate distortion function of a Gaussian process with a weighted square error criterion," *IEEE Trans. Inform. Theory*, vol. 14, pp. 506-508, 1968.
[21] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1971.
[22] R. M. Gray, *Source Coding Theory.* Boston, MA: Kluwer, 1990.
[23] A. A. M. L. Bruekers and A. W. M. van den Enden, "New networks for perfect inversion and perfect reconstruction," *IEEE J. Select. Areas Commun.*, this issue, pp. 129-137.
[24] P. P. Vaidyanathan, "Multirate digital filters," *Proc. IEEE*, vol. 78, no. 1, pp. 56-93, 1990.
[25] J. H. Rothweiler, "Polyphase quadrature filters, a new subband coding technique," in *Proc. ICASSP'83*, Boston, MA, 1983, pp. 1980-1983.
[26] P. L. Chu, "Quadrature mirror filter design for an arbitrary number of equal band-width channels," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 1, pp. 203-218, 1985.
[27] B. Scharf, "Critical bands," in *Foundations in Modern Auditory Theory.* New York: Academic, 1970, pp. 150-202.



Raymond N. J. Veldhuis (M'85) was born in The Hague, The Netherlands, on April 8, 1955. He received the Ingenieur degree in electronics from Twente University, Enschede, The Netherlands, in 1981, and the Ph.D. degree from the University of Nijmegen, Nijmegen, The Netherlands, in 1988.

In 1982 he joined Philips Research Laboratories, Eindhoven, The Netherlands, where he has been working in various fields of digital signal processing, among which sample restoration for digital audio, video, and speech signals, and source coding for digital audio. He has been involved in the development of source coding algorithms for DCC and MUSICAM.