PERCEPTUAL AUDIO CODING USING N-CHANNEL LATTICE VECTOR QUANTIZATION

Jan Østergaard, Omar Niamut, Jesper Jensen and Richard Heusdens

Information and Communication Theory Group,
Delft University of Technology,
2628 CD Delft, The Netherlands,
email: {j.ostergaard,o.niamut,j.jensen,r.heusdens}@ewi.tudelft.nl

ABSTRACT

We consider the problem of reliable distribution of audio over packet-switched networks. We make use of multiple-description coding combined with transform coding in order to obtain robustness towards packet losses. Previous approaches to this problem were restricted to the case of only two descriptions. In this work we use n-channel multiple-description lattice vector quantizers (MD-LVQs), which allow for the possibility of using more than two descriptions. For a given packet-loss probability we find the number of descriptions and the bit allocation between transform coefficients which minimizes a perceptual distortion measure subject to an entropy constraint. The optimal quantizers are presented in closed form, thus avoiding any iterative quantizer design procedures. The theoretical results are verified with numerical computer simulations using audio signals and it is shown that in environments with excessive packet losses it is advantageous to use more than two descriptions. We verify in subjective listening tests that using more than two descriptions lead to signals of perceptually higher quality.

1. INTRODUCTION

Services such as Voice over IP, audio streaming and video conferencing usually ask for high bandwidth, low delay and low packet-loss rates in order to deliver tolerable quality for the end users. However, the heterogeneous communication infrastructure of today's packet-switched networks do not provide a guaranteed performance in terms of e.g. bandwidth or delay and therefore the desired quality of service may not be achieved.

In this work we propose to mitigate the effect that unreliable channels have on the quality of audio streaming by use of multiple-description coding (MDC) [1]. The idea behind MDC is to create separate descriptions individually capable of reproducing a source to a specified accuracy and when combined being able to refine each other. Hence, in order to combat the effect of excessive audio packet losses we choose to transmit multiple audio packets simultaneously.

Traditionally, MDC schemes employ only two descriptions [2–5]. However, MDC schemes capable of exploiting more than two descriptions have recently been proposed [6–9]. The schemes presented in [6–8] consider the symmetric case where the side distortions and the side rates are all equal. In the asymmetric case [9] the side distortions and side rates are allowed to be unequal. In this work we will use the MDC schemes presented in [7,8] and as such we will focus on the symmetric case but our results can easily be extended to the asymmetric case by replacing the underlying MDC schemes by the ones proposed in [9].

This research is supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the ministry of Economics Affairs.

State-of-the-art audio coding schemes perform time-frequency analysis of the source signal which makes it possible to exploit perceptual models in both the time and the frequency domain in order to discard perceptual irrelevant information. This is done in e.g. MP3 [10], MPEG-2 advanced audio coding (AAC) [11] and Lucent PAC [12]. The time-frequency analysis is often done by a transform coder which is applied to blocks of the input signal. A common approach is to use the modified discrete cosine transform (MDCT) [13] as was done in e.g. MPEG-2 AAC and Lucent PAC. In this paper we propose to combine the MDCT with an n-channel MDC scheme based on multiple-description lattice vector quantizers (MD-LVQs) [7,8] in order to obtain a perceptual transform coder for audio which is robust to packet losses.

Multiple-description coding of audio has to the best of the authors knowledge so far only considered two descriptions [4,5]. In this work we propose a scheme that is able to use an arbitrary number of descriptions without violating an allowable target entropy. We show how to distribute the bit budget among the MDCT coefficients and present closed-form expressions for the MD-LVQs which minimize a perceptual distortion measure given the packet-loss characteristics of the transmission channel.

2. TRANSFORM CODING FRAMEWORK

The MDCT is a so-called 50% overlapped block transform, i.e. a transform where samples from consecutive 50% overlapping segments are windowed and transformed. Let s denote a segment of length $2M$ of the input signal. The set of M transform coefficients x is then obtained by applying the direct MDCT, which is defined as [13]

$$x_k = \frac{1}{\sqrt{2M}} \sum_{n=0}^{2M-1} h_n s_n \cos\left(\frac{(2n+M+1)(2k+1)\pi}{4M}\right)$$

where $k = 0, \dots, M-1$ and $x_k \in \mathbb{R}$ and $h \in \mathbb{R}^{2M}$ is an appropriate analysis window.

On each segment s a psycho-acoustic analysis is performed which leads to a masking curve that describes thresholds in the frequency domain below which distortions are inaudible. In our work we derive a set of perceptual weighting coefficients μ from the psycho-acoustic model presented in [14]. More specifically, we define μ as the inverse of the masking curve given in [14] evaluated at the center frequencies of the MDCT basis functions. With this we define a perceptual distortion measure in the MDCT domain between x and a quantized version $\hat{x}$ of x to be the single-letter distortion measure given by

$$d_\pi(x, \hat{x}) \triangleq \sum_{k=0}^{M-1} \mu_k |x_k - \hat{x}_k|^2. \quad (1)$$

We can rewrite (1) as $(y - \hat{y})^T (y - \hat{y})$, where $y_k = x_k\sqrt{\mu_k}, k = 0, \dots, M-1$. This representation has the advantage that the y_k's are all approximately equally sensitive to distortion after being "flattened" by the masking curve, c.f. [15]. In addition it is easy to show that the distortion measure on y and $\hat{y}$ approximates an ℓ_2 distortion measure for which it is known that lattice vector quantizers are among the set of optimal quantizers [16].

3. MULTIPLE-DESCRIPTION LATTICE VECTOR QUANTIZATION

3.1. Introduction to MD-LVQ

Let $Y \in \mathbb{R}^M$ be a random vector representing normalized MDCT coefficients and let y be a realization of Y. An MD-LVQ consists of one central lattice $\Lambda_c \subset \mathbb{R}^L$, where L is the lattice dimension[1] and K sublattices $\Lambda_i \subseteq \Lambda_c, i = 0, \dots, K-1$. The central lattice is also called the central quantizer and the sublattices are usually referred to as side quantizers. With MD-LVQs a source vector y is first quantized to the nearest reconstruction vector λ_c in the central lattice Λ_c. Hereafter follows index assignments (mappings), which uniquely map all λ_c's to reconstruction points in each of the sublattices Λ_i. This mapping is done through a labeling function $\alpha(\lambda_c) = (\lambda_0, \lambda_1, \dots, \lambda_{K-1})$.

At the receiving side y is reconstructed to a quality that is determined by the number of received descriptions. If no descriptions are received we reconstruct using the expected value, $E[Y]$, and if all K descriptions are received we reconstruct using the inverse map α^{-1}, hence obtaining the quality of the central quantizer. In all other cases, we reconstruct to the average of the received descriptions. There are in general several ways of receiving κ out of K descriptions. Let $\mathcal{L}$ denote an index set consisting of all possible κ combinations out of $\{0, \dots, K-1\}$. We denote an element of $\mathcal{L}$ by $l = \{l_0, \dots, l_{\kappa-1}\} \in \mathcal{L}$. Upon reception of any $\kappa < K$ descriptions we reconstruct to $\hat{y}$ using $\hat{y} = \frac{1}{\kappa}\sum_{i\in l}\lambda_i$.

3.2. Rate and distortion results of MD-LVQ

The results of this subsection are presented in detail in [7, 8]. The entropy R_c of the central quantizer is given by

$$R_c \approx h(Y) - \frac{1}{L}\log_2(\nu),$$

where $h(Y)$ is the componentwise differential entropy of Y and ν is the volume of a fundamental region of Λ_c. The entropies of the side quantizers are given by

$$R_i = R_c - \frac{1}{L}\log_2(N_i),$$

where $N_i = [\Lambda_i : \Lambda_c]$ is the index value of the sublattice Λ_i. Since we consider the symmetric case we have $N_i = N$ and $R_i = R$ for all i.

We will use a simple network model where the packet-loss probabilities are equal for all packets and in addition we assume the packet-loss probabilities to be independent. With this we can write the expected distortion based on the packet-loss probability p as

$$\begin{aligned} d &= \hat{K}_1 G(\Lambda_c) 2^{2(h(Y)-R_c)} \\ &+ \hat{K}_2 \psi_L^2 G(S_L) 2^{2(h(Y)-R_c)} 2^{\frac{2K}{K-1}(R_c-R)} + p^K \frac{1}{L} E[\|Y\|^2], \end{aligned} \quad (2)$$

where $\hat{K}_1 = 1 - p^K$ and $\hat{K}_2$ is given by

$$\hat{K}_2 = \sum_{\kappa=1}^{K} \binom{K}{\kappa} p^{K-\kappa}(1-p)^{\kappa}\frac{K-\kappa}{2\kappa K}. \quad (3)$$

[1] At this point we assume $L = M$. A restriction that will be removed later.

The constants $G(\Lambda_c)$ and $G(S_L)$ in (2) describe the dimensionless normalized second moment of inertia of respectively the central lattice Λ_c and an L-dimensional sphere [16]. The constant ψ_L is the dimensionless expansion factor, see [7, 8] for details.

The scaling of the central quantizer is given by ν and depends upon the source and channel characteristic. The optimal ν is given by

$$\nu = 2^{L(h(Y)-R)} \left(\frac{1}{K-1}\frac{\hat{K}_2}{\hat{K}_1}\frac{G(S_L)}{G(\Lambda_c)}\psi_L^2\right)^{\frac{L(K-1)}{2K}}, \quad (4)$$

and the optimal redundancy N is, for a fixed K, independent of the sublattices as well as source and target entropies, that is

$$N = \left((K-1)\frac{\hat{K}_1}{\hat{K}_2}\frac{G(\Lambda_c)}{G(S_L)}\frac{1}{\psi_L^2}\right)^{\frac{L(K-1)}{2K}}. \quad (5)$$

3.3. Optimal bit distribution

Each segment s leads to M MDCT coefficients which we vector quantize using an MD-LVQ. Since the number of coefficients in the MDCT is quite large, e.g. $M = 1024$ in our case, it is necessary to split the sequence of M coefficients into smaller vectors to make the quantization problem practically feasible. Any small number of coefficients can be combined and jointly quantized. For example if the set of M coefficients is split into M' bands (vectors) of length L_k where $k = 0, \dots, M'-1$ it can be seen from (2) that the distortion of band k is given by

$$\begin{aligned} d_k &= \hat{K}_1 G(\Lambda_k) 2^{2(h(Y_k)-R_{c_k})} + \hat{K}_2 \psi_{L_k}^2 G(S_k) 2^{2(h(Y_k)-R_{c_k})} \\ &\times 2^{\frac{2K}{K-1}(R_{c_k}-R_k)} + p^K \frac{1}{L} E[\|Y_k\|^2], \end{aligned} \quad (6)$$

where we allow the quantizers Λ_k to vary among the M' bands. The number of packets K is fixed for all the M' bands in a given segment, but may vary from segment to segment. For a given target entropy R^* we need to find the individual entropies R_k for the M' bands, such that $\sum R_k = R^*/K$ and in addition we need to find the entropies R_{c_k} of the central quantizers. For simplicity we assume in the following that the M' bands are of equal dimension L and that geometrically similar central lattices Λ_c are to be used. Furthermore, Eqs. (4) and (5) hold for any bit distribution, hence we may insert (4) and (5) into (6) which leads to individual distortions given by

$$d_k = a_0 2^{h(Y_k)-R_k} + p^K \frac{1}{L} E[\|Y_k\|^2],$$

where a_0 is independent of k and given by

$$a_0 = \hat{K}_1 G(\Lambda_c) \left(\frac{1}{K-1}\frac{\hat{K}_2}{\hat{K}_1}\frac{G(S_L)}{G(\Lambda_c)}\psi_L^2\right)^{\frac{K-1}{K}}.$$

In order to find the optimal bit distribution among the M' bands subject to the entropy constraint $\sum R_k = R^*/K$ we take the common approach of turning the constrained optimization problem into an unconstrained problem by introducing a Lagrangian cost functional of the form

$$J = \sum_{k=0}^{M'-1} d_k + \lambda \sum_{k=0}^{M'-1} R_k. \quad (7)$$

Differentiating (7) w.r.t. R_k leads to

$$\frac{\partial J}{\partial R_k} = -2\ln(2) a_0 2^{2(h(Y_k)-R_k)} + \lambda. \quad (8)$$

After equating (8) to zero and solving for R_k we get

$$R_k = \frac{1}{2}\log_2\left(\frac{\lambda}{2\ln(2)a_0}\right) + h(Y_k). \quad (9)$$

In order to eliminate λ we invoke the sum-rate constraint $\sum R_k = R^*/K$ and rewrite (9) as

$$R^*/K = \frac{1}{2}\log_2(\lambda)M' - \frac{1}{2}M'\log_2(2\ln(2)a_0) + \sum_{k=0}^{M'-1} h(Y_k),$$

from which we obtain

$$\lambda = 2\ln(2)a_0 2^{\frac{2}{M'}(R^*/K - \sum h(Y_k))}. \tag{10}$$

We can now eliminate λ by inserting (10) into (9), that is

$$R_k = \frac{R^*/K - \sum h(Y_k)}{M'} + h(Y_k). \tag{11}$$

With the simple Lagrangian approach taken here there is no guarantee that the entropies R_k given by (11) are all non-negative. It is possible to extend the Lagrangian cost functional (7) by M' additional Lagrangian weights in order to obtain M' inequality constraints making sure that $R_k \geq 0$ in addition to the single equality constraint $\sum R_k = R^*/K$. However, in this particular case this approach does not appear to lead to a closed-form expression for the individual entropies R_k. It is not possible either to simply just setting negative entropies equal to zero since this will most likely violate the constraint $\sum R_k = R^*/K$. Instead we propose a procedure where we begin by considering all M' bands and then one-by-one eliminate bands having negative entropies. Among the bands getting assigned a negative entropy, we find the one having the largest negative entropy and exclude that one from the optimization process. This procedure continues until all entropies are positive or zero as shown below.

1. $\mathcal{I} = \{0, \dots, M'-1\}$
2. $c = \frac{R^*/K - h}{|\mathcal{I}|}$, where $h = \sum_{k\in\mathcal{I}} h(Y_k)$
3. $\mathcal{R} = \{R_k : R_k = c + h(Y_k)$ and $R_k < 0, k \in \mathcal{I}\}$
4. If $|\mathcal{R}| > 0$ then goto 2 and set $\mathcal{I} := \mathcal{I}\backslash j$, where $R_j \leq R_k, \forall k \in \mathcal{I}$
5. $R_k = \begin{cases} c + h(Y_k) & k \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases}$

Table 1. Bit allocation algorithm.

The motivation for this approach is that ultimately we would like the contribution of each band to the total distortion to be equal, since they are all approximately equally sensitive to distortion. However, the MDCT coefficients in some bands have variances which are smaller than the average distortion, hence assigning zero bits to these bands leads to distortions which are lower than the average distortion over all bands. Therefore, the bit budget should only be distributed among the higher variance components.

4. ROBUST PERCEPTUAL TRANSFORM CODER

4.1. Encoder

We will use the transform coder presented in Section 2. Each segment is encoded into K descriptions independent of previous segments in order to avoid that the decoder is unable to successfully reconstruct due to previous description losses.

As discussed in Section 3.3 it is infeasible to jointly encode the entire set of M normalized MDCT coefficients and instead we split y into disjoint subsets. Based on the differential entropies of y and the target entropy R^* we find the individual entropies $R_k, k = 0, \dots, M'-1$ by using the algorithm described in Table 1.

The entropy R_k describes the total entropy assigned to the kth subset of bands. If the number of descriptions is K then each side quantizer operates at an entropy of R_k/K bits. Knowledge of R_k, the differential entropy $h(Y_k)$, the number of descriptions K and the packet-loss probability p makes it possible to find the scaling factors ν_k and N_k of the central and side quantizers, respectively by use of (4) and (5). This in turn completely specifies an MD-LVQ having K descriptions. Each normalized vector of MDCT coefficients y_k is then first quantized with the central quantizer $Q_k(y_k) = \lambda_{c_k}$ after which an index assignment $\alpha(\lambda_{c_k}) = \{\lambda_{0_k}, \dots, \lambda_{K_k-1}\}$ is applied in order to find the codewords of the side quantizers. The codewords of the side quantizers are losslessly encoded and put into K individual packets. Each packet then contains M' encoded codewords.

In order to be able to reconstruct the quantized MDCT coefficients it is required that the perceptual weight μ is included in all K packets. In [17] it was shown that μ can be effectively encoded at 4 kbps. In this work we assume that μ can be encoded at 4 kbps, hence if the total target entropy is $R^* = 96$ kbps and two packets are to be used ($K = 2$), the entropy we actually use for the MD-LVQ is then only 88 kbps.

4.2. Decoder

At the receiving side an estimate $\hat{y}$ of the normalized MDCT spectrum is first obtained by simply taking the average of the received descriptions, i.e. $\hat{y}_k = \frac{1}{\kappa}\sum_{i\in l}\lambda_{i_k}$, where l denotes the indices of the received descriptions and $\kappa = |l|$. This estimate is then denormalized in order to obtain $\hat{x}$, i.e. $\hat{x}_k = \hat{y}_k/\sqrt{\mu_k}$. Finally the inverse MDCT is applied in order to obtain an approximation $\hat{s}$ of s.

5. RESULTS

In this section we compare numerical simulations with theoretical results and in addition we show the results of a subjective listening test. For both tests we use three audio clips of different genres (jazz, German male speech and pop) each having a duration between 10 and 15 sec. and a sampling frequency of 48 kHz. We set the target entropy to 96 kbps which corresponds to 2 bit/dim. For simplicity we assume that the sources are stationary processes so that we can measure the statistics for each vector of MDCT coefficients upfront. However, since audio signals in general have time varying statistics we expect that it will be possible to reduce the bit rate by proper adaptation to the source. Since for this particular test we are merely interested in the performance of the proposed audio coder with a varying number of descriptions we will not address the issue of efficient entropy coding but simply assume that the quantized variables can be lossless encoded arbitrarily close to their discrete entropies. Table 2 shows the discrete entropies of the quantized normalized MDCT coefficients for the three test fragments.

	$K=2$ kbps	$K=2$ bit/dim	$K=3$ kbps	$K=3$ bit/dim	$K=4$ kbps	$K=4$ bit/dim
Jazz	96.22	1.00	97.09	0.67	96.87	0.51
Speech	93.48	0.98	96.00	0.67	96.47	0.50
Pop	93.76	0.98	95.38	0.66	95.60	0.50

Table 2. Discrete entropies. The columns showing [bit/dim] are expressed as entropies per description.

Because of the short duration of the test fragments the resulting expected distortions depend upon the realizations of the packet loss patterns. This phenomenon has been noted by other authors, c.f. [4]. We therefore decided to average the distortion results over three different loss patterns. The theoretical and numerical obtained expected distortions for the jazz signal are shown in Fig. 1.

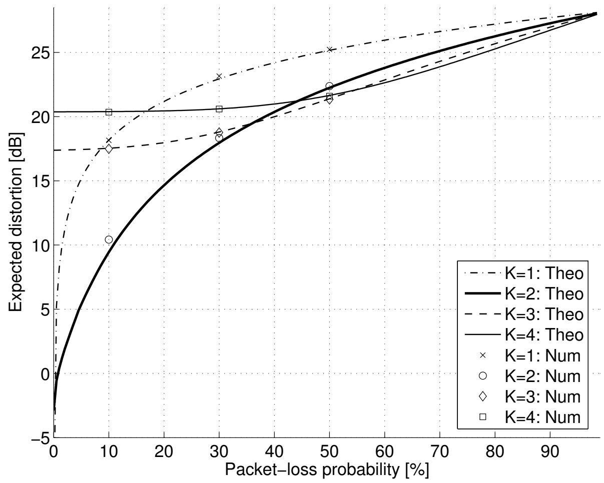

Fig. 1. Expected distortion as a function of packet-loss probability.

As can be seen in Fig. 1 the expected distortions depend not only on the packet-loss rate but also upon the number of descriptions. At high packet-loss rates it is advantageous to use a higher number of packets.

To verify these findings we performed an additional subjective comparison (MUSHRA [18]) test, see Fig. 2. The packet-loss rates in the test are $p = 0.1$, $p = 0.3$ and $p = 0.5$ and we have averaged results over nine listeners and over the three different test fragments. At each packet-loss rate the original signals were encoded using $K = 1, 2, 3$ and 4 descriptions. Also included in each test were the hidden reference and two anchor signals (3.5 kHz and 7 kHz lowpass filtered signals). The circles in Fig. 2 denote mean values and the bars describe 95% confidence intervals. Notice that for $p = 0.3$ and $p = 0.5$ there is a significant preference for using more than two descriptions.

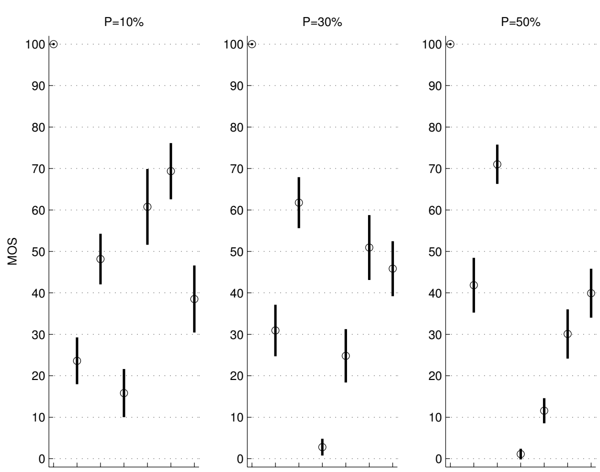

Fig. 2. MUSHRA test results where the seven signals appear in the following order: Hidden ref., 3.5 kHz, 7 kHz, $K = 1, 2, 3$ and 4.

6. REFERENCES

[1] V. K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Processing Mag.*, pp. 74 – 93, Sep. 2001.

[2] V. A. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Trans. Inform. Th.*, vol. 39, no. 3, pp. 821 – 834, May 1993.

[3] V. A. Vaishampayan, N. J. A. Sloane, and S. D. Servetto, "Multiple-description vector quantization with lattice codebooks: Design and analysis," *IEEE Trans. Inform. Th.*, vol. 47, no. 5, pp. 1718 – 1734, July 2001.

[4] R. Arean, J. Kovačević, and V. K. Goyal, "Multiple description perceptual audio coding with correlating transform," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 2, pp. 140 – 145, March 2000.

[5] G. Schuller, J. Kovačević, F. Masson, and V. K. Goyal, "Robust low-delay audio coding using multiple descriptions," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 1014 – 1024, September 2005.

[6] S. S. Pradhan, R. Puri, and K. Ramchandran, "n-channel symmetric multiple descriptions–part I: (n, k) source-channel erasure codes," *IEEE Trans. Inform. Th.*, vol. 50, no. 1, pp. 47 – 61, Jan. 2004.

[7] J. Østergaard, J. Jensen, and R. Heusdens, "n-channel symmetric multiple-description lattice vector quantization," in *Proc. Data Compr. Conf.*, March 2005, pp. 378 – 387.

[8] J. Østergaard, J. Jensen, and R. Heusdens, "n-channel entropy-constrained multiple-description lattice vector quantization," *Submitted to IEEE Trans. Inform. Th.*

[9] J. Østergaard, R. Heusdens, and J. Jensen, "n-channel asymmetric multiple-description lattice vector quantization," in *Proc. IEEE Int. Symp. on Inform. Th.*, September 2005, pp. 1793 – 1797.

[10] International Standard ISO/IEC 11172-3 (MPEG), "Information technology - coding of moving pictures and associated audio for digital storage media at up to about 1.5 mbit/s. part 3: Audio," 1993.

[11] International Standard ISO/IEC 13818-7 (MPEG), "Information technology - generic coding of moving pictures and associated audio, part 7: Advanced audio coding," 1997.

[12] D. Sinha, J. D. Johnston, S. Dorward, and S. Quackenbush, "The perceptual audio coder (PAC)," in *The digital signal processing handbook*. New York: IEEE Press, 1998.

[13] J.P. Princen, A.W. Johnson, and A.B. Bradley, "Subband/transform coding using filter bank designs based on time domain aliasing cancellation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, April 1987, pp. 2161 – 2164.

[14] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psychoacoustical masking model for audio coding applications," in *Proc. of the 2002 Int. Conf. on Acoustics, Speech and Signal Processing*, May 2002, pp. 1805–1808.

[15] B. Edler and G. Schuller, "Audio coding using a psychoacoustic pre- and post-filter," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, 2000, vol. II, pp. 881 – 884.

[16] J. H. Conway and N. J. A. Sloane, *Sphere packings, Lattices and Groups*, Springer, 3rd edition, 1999.

[17] O. Niemeyer and B. Edler, "Efficient coding of excitation patterns combined with a transform audio coder," in *Proc. of the 118th AES Conv.*, Barcelona, Spain, May 2005.

[18] "Methods for the subjective assessment of intermediate quality levels of coding systems," ITU-R BS. 1534-1, 2003.