MULTIPLE DESCRIPTION QUANTIZATION OF SINUSOIDAL PARAMETERS

Morten Holm Larsen, Mads Græsbøll Christensen, and Søren Holdt Jensen

Department of Electronic Systems
Aalborg University, Denmark
{mhl,mgc,shj}@es.aau.dk

ABSTRACT

A new scheme for sinusoidal audio coding named multiple description spherical trellis-coded quantization is proposed and analytic expressions for the point densities and expected distortion of the quantizers are derived based on a high-resolution assumption. The proposed quantizers are of variable dimension, i.e., sinusoids can be quantized jointly for each audio segment whereby a lower distortion is achieved. The quantizers are designed to minimize a perceptual distortion measure subject to an entropy constraint for a given packet-loss probability. In experiments, the performance of the quantizers is compared to the corresponding single description spherical quantizer and associated bounds are found to increase robustness towards packet-losses.

Index Terms— Quantization, audio coding, robustness.

1. INTRODUCTION

Parametric audio coding is based on the notion that most audio signals can be efficiently described by a few physically or perceptually meaningful parameters. Perhaps the most common incarnation is sinusoidal coding where the individual audio segments are modeled as sums of sinusoids with each of these being characterized by an amplitude, a phase, and a frequency that combine to form a point in a spherical coordinate system. For each segment of audio, the task is to find the parameters best describing the segment and to quantize these parameters, whereby transmission over channels of limited capacity is facilitated. Various computationally efficient ways of finding the parameters that minimize a perceptual distortion measure exist (see, e.g., [1]). Also, the question of optimal quantization of these parameters has been addressed recently. The so-called polar and spherical quantizers of [2, 3] have proven successful in terms of achieved quality and computational complexity by quantization of the parameters of each sinusoid independently. The quantizers were designed to minimize a perceptual distortion measure subject to an entropy constrain based on a high-resolution assumption, i.e., a high number of bits per sinusoid, whereby analytic expressions for the point densities of the quantizers were derived. In [4], the spherical quantizers of [3] were improved by joint quantization of the parameters of a variable number of sinusoids. Under a high-resolution assumption, the optimal point densities of the proposed quantization scheme, named spherical trellis-coded quantization (STCQ), were derived for a given entropy. In services such as speech coding or audio streaming over unreliable networks like the Internet, the transmitted audio parameters should be protected to compensate for packet-losses. One method that aims at doing this is multiple description coding where several complementary coarse descriptions of the audio signal are constructed and transmitted whereby graceful degradation is achieved when packets are lost. Multiple description coding has recently been applied to audio in the form of transform coding [5, 6] and low-delay coding using pre- and post-filtering [7]. A limitation of the quantizer used in the latter paper is that the dimension of the vector must be fixed. Therefore, it cannot readily be applied to the problem of joint quantization of sinusoidal parameters. The multiple description trellis-coded quantizer (MDTCQ) of [8] can, on the other hand, handle variable dimensions but requires training for a particular combination of entropy constraint and packet-loss probability.

In this paper, we extend the spherical quantizers of [3] to multiple descriptions by proposing a quantization scheme, named multiple description spherical trellis-coded quantization (MDSTCQ). Based on high-resolution theory, we derive analytic expressions for the expected distortion and point densities for a given target entropy and packet-loss probability. The MDSTCQ is based on a new quantization scheme named modified multiple description trellis-coded quantization (MMDTCQ) that can be analytically designed from its point density given a packet-loss probability.

2. PROBLEM STATEMENT

We start this section by introducing the mathematical problem of robust quantization in parametric audio coding based on a perceptually relevant distortion measure. Let the audio signal x at sample time n be represented as $x(n) \approx \sum_{l=1}^{L} a_l \sin(\nu_l n + \phi_l)$, where L is the number of sinusoidal components and a_l, ϕ_l, ν_l are the amplitude, phase and frequency of the l'th component, respectively, with $a_l \geq 0$ and $\phi_l, \nu_l \in [0, 2\pi)$. The quantization distortion consists of the contributions from the individual components and the cross-terms between the components. Assuming a sufficiently large window length W or statistical independence between components, the total expected distortion can be approximated as the sum over the L expected distortions for the individual components, denoted $E[D]$, with $E[\cdot]$ being the expectation operator. Therefore, we will in the rest of this paper be concerned with the quantization of a single set of parameters (a, ϕ, ν) thus ignoring the subscript l. The present work is

M. G. Christensen is supported by the Parametric Audio Processing project, Danish Research Council for Technology and Production Sciences grant no. 274–06–0521.

1-4244-1484-9/08/$25.00

based on a perceptual distortion measure [9] defined as

$$D = \frac{1}{2\pi}\int_0^{2\pi} \mu_{x(a,\phi,\nu)}(\omega)|E(\omega)|^2 d\omega, \quad (1)$$

with $E(\omega)$ denoting the Fourier transform of the windowed error, i.e., $E(\omega) = \sum_{n=n_0}^{n_0+W-1} w(n)(x(n) - \tilde{x}(n))e^{-j\omega n}$ where $w(n)$ is the window, $\mu_{x(a,\phi,\nu)}(\omega)$ is the perceptual weighting function, which is calculated from the audio signal x parametrized by (a, ϕ, ν). Furthermore, $\tilde{x}$ is the reconstructed audio signal based on the quantized parameters $(\tilde{a}, \tilde{\phi}, \tilde{\nu})$. Next, we introduce the quantization errors $\epsilon_a = a - \tilde{a}$, $\epsilon_\phi = \phi - \tilde{\phi}$, $\epsilon_\nu = \nu - \tilde{\nu}$, and the constant $\|w\|^2 = \sum_{n=n_0}^{n_0+W-1} w^2(n)$. Then, assuming a large W, high-resolution, and a smooth masking curve, the perceptual distortion can be approximated as $D \approx \frac{\mu_{x(a,\phi,\nu)}}{2}(\|w\|^2(a^2+\tilde{a}^2) - 2a\tilde{a} \times \sum_{n=n_0}^{n_0+W-1} w^2(n)\cos(\epsilon_\nu n + \epsilon_\phi))$. Similarly to [2, 3], we assume the perceptual weighting function to be quantized and transmitted as side information. To the best of our knowledge, the problem of joint quantization of the perceptual weighting function and the sinusoidal parameters remains unsolved and we will defer from any further discussion of this. Setting $n_0 = -\frac{W}{2}$, and using a Taylor expansion, the distortion can be shown to be

$$D \approx \frac{\mu_{x(a,\phi,\nu)}}{2\|w\|^{-2}}\left(\epsilon_a^2 + a\tilde{a}\left(\epsilon_\phi^2 + \epsilon_\nu^2\sigma^2\right)\right), \quad (2)$$

with $\sigma^2 = \frac{1}{\|w\|^2}\sum_{n=-W/2}^{W/2-1} w^2(n)n^2$. We observe from (2) that the amplitude, phase and frequency can be quantized independently using the l_2-norm by assuming a high-resolution, i.e., $a\tilde{a} \approx a^2$. This provides inspiration to the proposed multiple description spherical trellis-coded quantization (MDSTCQ) scheme consisting of three multiple description quantizers, one for each of the parameters a, ϕ and ν. We here focus on the common case of two descriptions. In this case, a low quality description is obtained when only one description is received. The resulting distortion, called the side distortion, is denoted as D_s, with $s = \{1, 2\}$. The low quality reconstructions of (a, ϕ, ν) is written as $(\tilde{a}_s, \tilde{\phi}_s, \tilde{\nu}_s)$. When both descriptions are received, the resulting distortion is referred to as the central distortion D_c and is based on the reconstruction point $(\tilde{a}_c, \tilde{\phi}_c, \tilde{\nu}_c)$. The aim of this work is to protect the sinusoidal parameters transmitted over a packet erasure channel where packets are dropped independently with probability p. Assuming a balanced side distortion, i.e., $E[D_1] = E[D_2]$, the average distortion in such a network is

$$E[D] = (1-p)^2 E[D_c] + 2p(1-p)E[D_s] + p^2 E[x^2], \quad (3)$$

where $E[x^2]$ is the variance of the audio signal.

3. THE PROPOSED QUANTIZER

We now proceed to propose a modified multiple description trellis-coded quantizer (MMDTCQ) inspired by the two-stage coding scheme in [10]. The first stage is to quantize the input signal y using two uniform side quantizers, Q_1 and Q_2 that are offset to each other and have step-sizes equal to the reciprocal value of the point density, g_y, as shown in Fig. 1. In the second stage, we perform joint quantization using the joint Voronoi region, which is half the size of the Voronoi region of the side quantizers. Here, N reconstruction points are

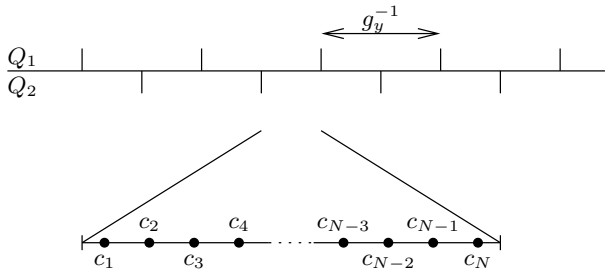

Fig. 1. Structure of the MMDTCQ with the vertical lines being the Voronoi regions of the two side quantizers Q_s and dots the refined reconstruction points.

sorted as $c_1 < c_2 < \cdots < c_N$ and partitioned into four subsets $C_q = \{c_q, c_{q+4}, \cdots, c_{N-4+q}\}$, where $q = \{1, 2, 3, 4\}$. In trellis-coded quantization, the transitions in the trellis specify C_q by one bit, and $\log_2(N/4)$ bits are used to specify the quantization index within C_q. Since this refined trellis information is only utilized when both descriptions are received, it is easily split into the two descriptions whereby the otherwise difficult index assignment is solved. Asumming high-resolution such that the probability density function (pdf) inside the Voronoi regions V_i of the side quantizers Q_s is uniform, the mean square-error distortion for each side quantizer can be written as $E[D_s] = \sum_{i\in I}\int_{V_i} f_Y(y)\epsilon_y^2 dy \approx g_y^{-2}/12$ with $f_Y(y)$ being the source pdf and $\epsilon_y = y - \tilde{y}$. An exact expression for the expected distortion does not, however, exist for the TCQ. Therefore, we will employ the approximation proposed in [11] where the expected distortion is written as the distortion for a uniform quantizer corrected by a factor Γ, which depends on the trellis structure, number of states and dimension, that has been determined numerically in [4]. Also, assuming high-resolution the central distortion can be written as $E[D_0] \approx \frac{\Gamma}{(2N)^2}E[D_s]$. The entropy of a MMDTCQ per description is given by $H \approx h(Y) + \int f_Y(y)\log_2(g_y)dy + \frac{1}{2}\log_2(\frac{N_y}{2})$, where $h(Y)$ is the differential entropy of the source Y.

We will now introduce the details of the proposed MDSTCQ coding scheme consisting of three MMDTCQs, one for each of a, ϕ and ν. In deriving the optimal MDSTCQ design, we need expressions for the quantization point densities and the number N for the three MMDTCQs. To obtain these, we first introduce the joint pdf $f(A, \phi, \nu)$ whereafter we can express the expected side distortion as

$$E[D_s] \approx \sum_{i_a\in I_A}\sum_{i_\phi\in I_\phi}\sum_{i_\nu\in I_\nu}\int_{V_{i_a}}\int_{V_{i_\phi}}\int_{V_{i_\nu}} f(A,\phi,\nu)\frac{1}{2}\mu_x(\tilde{\nu})$$
$$\|w\|^2\left(\epsilon_a^2 + a\tilde{a}_s\left(\epsilon_\phi^2 + \epsilon_\nu^2\sigma^2\right)\right)dad\phi d\nu \approx \frac{\|w\|^2}{24}\int_A\int_\phi\int_\nu$$
$$f(A,\phi,\nu)\mu_x(\tilde{\nu})\left(g_a^{-2} + \tilde{a}_s^2\left(g_\phi^{-2} + \sigma^2 g_\nu^{-2}\right)\right)dad\phi d\nu \quad (4)$$

by assuming that $\mu_{x(a,\phi,\nu)}$ is constant over the joint Voronoi region of a, ϕ, ν. We have assumed high-resolution such that $a\tilde{a}_s \approx \tilde{a}_s^2$ and the probability mass function of the reconstruction points, $\Pr(\tilde{a}_{i_a}, \tilde{\phi}_{i_\phi}, \tilde{\nu}_{i_\nu})$ can be found from the joint pdf as $f(\tilde{a}_{i_a}, \tilde{\phi}_{i_\phi}, \tilde{\nu}_{i_\nu})g_a^{-1}g_\phi^{-1}g_\nu^{-1}$ (see [3] for more details on this). Here, the quantization point densities for a, ϕ, ν are written as $g_{\{a,\phi,\nu\}}$, although they at this point still depend on a, ϕ, ν.

Similarly, we can express the expected central distortion as

$$E[D_0] \approx \frac{\|w\|^2}{96}\Gamma \int_A \int_\phi \int_\nu f_{A,\phi,\nu}\mu_x(\tilde{\nu}) \left(\frac{g_a^{-2}}{N_a^2} + \tilde{a}_0^2 \left(\frac{g_\phi^{-2}}{N_\phi^2} + \sigma^2 \frac{g_\nu^{-2}}{N_\nu^2} \right) \right) da d\phi d\nu, \quad (5)$$

where, for simplicity, we have assumed an equal trellis structure, $\Gamma = \Gamma_{\{a,\phi,\nu\}}$. Also, N_a N_ϕ, N_ν are the number of reconstruction points for the various refined quantizers. Next, assuming that the amplitude, phase and frequency are independent, we can write the entropy for each description as

$$H_s \approx h(A,\Phi,\Upsilon) - \frac{3}{2} + \iiint f(a,\phi,\nu)(\log_2(g_a g_\phi g_\nu) + \frac{1}{2}\log_2(N_a N_\phi N_\nu)) da d\phi d\nu, \quad (6)$$

with $h(A,\Phi,\Upsilon)$ being the differential entropy. To simplify the notation, we introduce $\tilde{H}_s = H_s - h(A,\Phi,\Upsilon)$ and write the cost function as $J = (1-p)^2 E[D_0] + 2p(1-p)E[D_s] + \lambda \tilde{H}$ with λ being the Lagrange multiplier. We note that there is the somewhat subtle problem that the distortions depend on $\tilde{a}_s$ and $\tilde{a}_0$. The point densitites also depend on the amplitude a and in [3] it was argued that the amplitude can be replaced by its reconstruction due to the high-resolution assumption. Similarly, we here use $\tilde{a}_s$ in lieu of $\tilde{a}_0$ (later we will evaluate the loss, if any, of doing this). We now minimize this cost function by taking the derivative with respect to g_a, g_ϕ, g_ν, N_a, N_ϕ and N_ν. This results in $N_a = N_\phi = N_\nu = \sqrt{\Gamma(1-p)/8p} \triangleq N$ and the following expressions for the point densities:

$$g_a^2 = \frac{p(1-p)\mu_x(\tilde{\nu})\|w\|^2}{12\lambda \log_2(e)}, g_\phi^2 = g_a^2 \tilde{a}_s, g_\nu^2 = g_a^2 \tilde{a}_s \sigma^2. \quad (7)$$

Next, we insert these equations and the expression for N into the definition of $\tilde{H}_s$ and obtain the optimal λ^* as

$$\lambda^* = \frac{Np(1-p)\|w\|^2}{12\log_2(e)} 2^{\frac{2}{3}(h(a,\phi,\nu) - H_s + \log_2(\sigma) + 2\varrho) + 1 + \psi}, \quad (8)$$

where $\varrho = \int f_A(a)\log_2(a) da$ with $f_A(a)$ being the pdf of a and $\psi = \iiint f(a,\phi,\nu)\log_2 \mu_{x(a,\phi,\nu)} da d\phi d\nu$. Using λ^* we get the optimal point densities:

$$g_a = \left(\frac{\mu_x(\tilde{\nu})}{2N} \right)^{\frac{1}{2}} 2^{\frac{1}{3}(H_s - h(a,\phi,\nu) - \log_2(\sigma) - 2\varrho) - \frac{\psi}{2}}, \quad (9)$$

$$g_\phi = \tilde{a}_s g_a \quad \text{and} \quad g_\nu = \sigma \tilde{a}_s g_a. \quad (10)$$

By inserting these into (4),(5) and (3), we can finally determine the expected distortion as

$$E[D] \approx \|w\|^2 \sqrt{\frac{\Gamma p}{8}} (1-p)^{\frac{3}{2}} 2^{\frac{2}{3}(h(a,\phi,\nu) - H_s + \log_2(\sigma) + 2\varrho) + \psi} + p^2 E[x^2]. \quad (11)$$

The point densities for ϕ and ν can be seen to be functions of $\tilde{a}_s$ (or $\tilde{a}_0$), i.e., the point densitities depend on the reconstructed amplitude. Since the fundamental notion of multiple description coding is to transmit complementary descriptions that combined lead to a reduced distortion, the reconstructed amplitudes in each of the descriptions will be different, i.e., $\tilde{a}_1 \neq \tilde{a}_2$, leading to phase and frequency quantizers having slightly different resolution for each description. To arrive at a feasible scheme, we encode ϕ_s and ν_s based on $\tilde{a}_s$ for each of the two descriptions with $s = \{1,2\}$. In the ideal case, the point densities for each description are equaivalent and the quantizers can be perfectly offset as illustrated in Fig. 1, in which case the reconstruction points of the joint description can be obtained as the mean of the reconstruction points of the two descriptions plus the contribution from the refined trellis-coded quantization, e.g., $\tilde{\phi}_0 = \frac{\tilde{\phi}_1 + \tilde{\phi}_2}{2} + \tilde{\phi}_{TCQ}$. In our case, however, the resolution of the two quantizers are slightly different, and we propose to deal with this in the following way: We consider the two descriptions as random variables from which we seek to estimate the mean. Due to the slightly different amplitudes $\tilde{a}_1$ and $\tilde{a}_2$, the two observations have been subjected to additive noise having different variances. Therefore, the mean can be obtained for the phase as $\tilde{\phi}_0 = \tilde{\phi}_1 \zeta_\phi + \tilde{\phi}_2(1-\zeta_\phi) + \tilde{\phi}_{TCQ}$, where $0 \leq \zeta_\phi \leq 1$ is a weight. Since the quantizers are uniform, the observation noise can be modeled as uniform random variables having a variance corresponding to their Voronoi regions and the optimal weight ζ_ϕ can then easily be determined from the point densities as $\zeta_\phi = g_{\phi,1}^2/(g_{\phi,1}^2 + g_{\phi,2}^2)$ and similarly for the frequency ν.

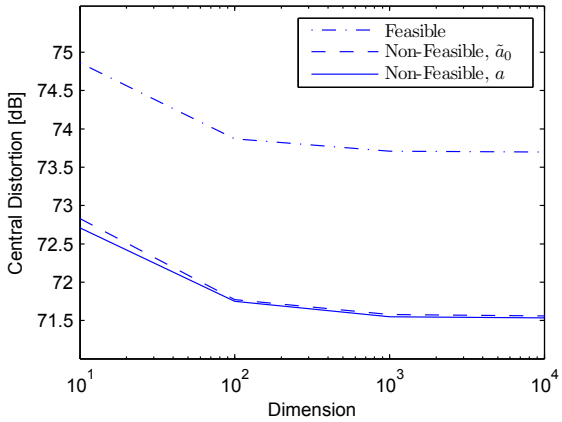

Fig. 2. Performance of the feasible MDSTCQ scheme compared to the two non-feasible MDSTCQ schemes where a and $\tilde{a}_0$ are used, respectively.

4. EXPERIMENTAL RESULTS

The two following experiments are based on synthetic audio, generated using a statistical model similar to that employed in [3]. Specifically, the amplitudes and frequencies are generated from a Rayleigh pdf, i.e., $f_Y(y) = \beta^{-2} y e^{-(y^2\beta^{-2}/2)}$, with $\beta = \{1000, 0.25\}$. The phase, on the other hand, is uniformly distributed in the interval $[0, 2\pi)$. We focus on the performance gain achieved by joint quantization and we will therefore, for simplicity, set the perceptual weighting function to one and use a rectangular window with $W = 1023$. We remark that the trellis is initialized in zero-state and that no side information is needed. In the first experiment, we will

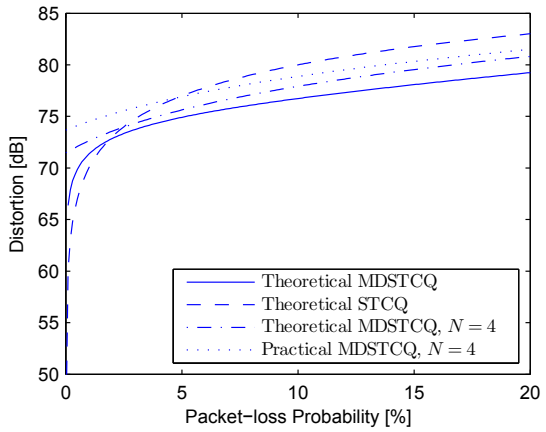

Fig. 3. The theoretical expected distortion for the MDSTCQ, the single desciption STCQ, and the practical MDSTCQ with $N = 4$.

investigate the impact of the number of jointly quantized sinusoids and the number of states in the trellis on the expected distortion. Furthermore, we will illustrate the impact of the estimation of ϕ_0 and ν_0 at the decoder. We generate 100,000 triplets (a,ϕ,ν), set the target entropy to 15 bit/sinusoid per description, using 256 state trellis with $N = 4$ and quantize the triplets. As explained, the optimal reconstruction of MDSTCQ is not feasible, but a feasible solution can be obtained by estimating $\tilde{\phi}_0$ and $\tilde{\nu}_0$. The performance of two non-feasible and the single feasible MDSTCQ schemes are shown in Fig. 2 for a range of dimension. From Fig. 2 it can be seen, that we gain about 1.2 dB when increasing the dimension from 10 to 10,000. Furthermore, it can be seen that there is a 2.2 dB gap between the non-feasible method and the feasible MDSTCQ method. We now compare the performance of the MDSTCQ to the theoretical expected distortion for both the MDSTCQ and the single description STCQ of [4]. As before, we generate 100,000 triplets and jointly quantize 1000 sinusoidal parameters using the MDSTCQ design and compare to a 256 state STCQ with 30 bit/sinusoid with one description for a packet-loss probability of p. The performance as a function of the packet-loss probability is shown in Fig. 3. From the figure, it can be seen that the theoretical bound of MDSTCQ is better than the theoretical bound of STCQ for a large range of packet-loss probabilities, except when packet-loss probabilities are close to zero. For the practical MDSTCQ, we limit ourselves to one design with $N = 4$ and compare it to the corresponding MDSTCQ theoretical bound. In the figure, a 2.2 dB gap between the practical and the theoretical performance can be seen for low p. This is, most likely, the same suboptimality observed in the previous experiment. However, we note that even this constant MDSTCQ design will outperform the single description STCQ for a large range of packet-loss probabilities.

5. CONCLUSION

We have proposed multiple description spherical trellis-coded quantization of sinusoids. The quantizers are suitable for parametric audio coding where the number of sinusoids may vary and for transmission over unreliable networks like the Internet. Under high-resolution assumptions we have derived analytical expressions for the optimal design and the expected perceptual distortion for a given target entropy and packet-loss probability. Experiments have shown significant performance improvements of the proposed scheme as the number of dimensions is increased. Furthermore, a significant performance gain compared to the single description spherical trellis-coded quantization scheme of [4] has been observed for a large range of packet-loss probabilities.

6. REFERENCES

[1] M. G. Christensen and S. H. Jensen, "On perceptual distortion minimization and nonlinear least-squares frequency estimation," *IEEE Trans. Audio, Speech and Lang. Processing*, vol. 14, no. 1, pp. 99–109, January 2006.

[2] R. Vafin, D. Prakash, and W.B. Kleijn, "On frequency quantization in sinusoidal audio coding," *IEEE Signal Processing Lett.*, vol. 12, no. 3, pp. 210–213, March 2005.

[3] Pim Korten, Jesper Jensen, and Richard Heusdens, "High-resolution spherical quantization of sinusoidal parameters," *IEEE Trans. Audio, Speech and Lang. Processing*, vol. 15, no. 3, pp. 966–981, March 2007.

[4] Morten Holm Larsen, Mads Græsbøll Christensen, and Søren Holdt Jensen, "Variable dimension trellis-coded quantization of sinusoidal parameters," *IEEE Signal Processing Lett.*, to appear.

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

[6] J. Østergaard, O. Niamut, J. Jensen, and R. Heusdens, "Perceptual audio coding using n-channel lattice vector quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2006, vol. 5, pp. 197–200.

[7] G. Schuller, J. Kovacevic, 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.

[8] V.A. Vaishampayan, A.R. Calderbank, and J.-C. Batllo, "On reducing granular distortion in multiple description quantization," in *Proc. IEEE Int. Symp. Information Theory*, 1998, p. 98.

[9] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," in *EURASIP J. on Applied Signal Processing*, 2005, vol. 9, pp. 1292–1304.

[10] Chao Tian and S.S. Hemami, "A new class of multiple description scalar quantizer and its application to image coding," *IEEE Signal Processing Lett.*, vol. 12, no. 4, pp. 329–332, April 2005.

[11] T.R. Fischer and M. Wang, "Entropy-constrained trellis-coded quantization," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 415–426, March 1992.