Variable Dimension Trellis-Coded Quantization of Sinusoidal Parameters

Larsen, Morten Holm; Christensen, Mads G.; Jensen, Søren Holdt

Published in:
IEEE Signal Processing Letters

Publication date:
2008

Document Version
Accepted author manuscript, peer reviewed version

Variable Dimension Trellis-Coded Quantization of Sinusoidal Parameters

Morten Holm Larsen, *Student Member, IEEE*, Mads Græsbøll Christensen, *Member, IEEE*, Søren Holdt Jensen, *Senior Member, IEEE*

***Abstract*—In this letter, we propose joint quantization of the parameters of a set of sinusoids based on the theory of trellis-coded quantization. A particular advantage of this approach is that it allows for joint quantization of a variable number of sinusoids, which is particularly relevant in variable rate parametric audio coding. Under high-resolution assumptions and based on a perceptually relevant distortion measure, we derive analytical expressions for the optimal design subject to an entropy constraint. Numerical experiments show a significant performance gain compared to optimal spherical quantization at the cost of a slight increase in computational complexity.**

I. Introduction

In sinusoidal coding, the audio signal is divided into a number of consecutive segments where each segment is modeled as a sum of sinusoids. The sinusoidal parameters, namely the amplitude, phase and frequency, are quantized and transmitted. Sinusoidal coding is suitable for audio coding at low bit-rates, where quantization of the sinusoidal parameters is of crucial importance for the audio quality. For a desired target bit-rate the joint optimum time segmentation (segment lengths), the distribution of sinusoidal components and quantization can be solved by dynamic programming [1]. To make the optimization feasible, the quantization and design must be low in terms of computational complexity, since distortion and rates have to be determined for each possible segment length and start position. Assuming high-resolution (high bit-rate per sinusoid) a simple analytical expression for the expected distortion and the spherical quantization (SQ) design can be derived, e.g. [2]–[4]. It is well known that vector quantization (VQ) outperforms scalar quantization and can approach the Shannon rate-distortion bound when the dimension is increased. For the problem at hand this can be done by joint quantization of sinusoidal parameters in a segment or even across segments. In [5], for example, each segment can be encoded using a variable number of sinusoids and using rate-distortion optimization, the jointly optimal distribution of sinusoids and segmentation are found. Assuming that each segment can be modeled using any number between 0 and 100 sinusoids and that the rate-distortion optimization is performed over super frames of about 0.5 s, we then get that, for a target bit-rate of 24 kbits/s and an average bit allocation of 20 bits/sinusoid, we can, in principle, perform joint quantization of 600 sinusoids without introducing further encoding delay. Usually, for a trained VQ using the K-means algorithm or lattice vector quantizers, all elements in the vector are quantized jointly, and the dimension must be *fixed* and known *a priori*. Therefore, these methods cannot readily be applied. The discipline of variable dimension quantization, it appears, has not gained much attention. Though, in [6], a method for variable dimension quantization was proposed and then in [7], this method was applied to the problem of quantization of sinusoidal parameters for speech coding. Specifically, this was done by a variable-to-fixed dimension transform followed by quantization.

In Trellis-Coded Quantization (TCQ) each transition (in the trellis) entails quantization of *one element* in the vector, and since the number of transitions is flexible, it can quantize vectors having a variable dimension. Furthermore, TCQ has been shown to be an effective technique both in terms of computations, storage and fidelity [8]–[10]. In this letter, we propose joint encoding of the sinusoidal parameters using a spherical trellis-coded quantization (STCQ). Under a high-resolution assumption, we derive analytical expressions of the expected distortion and quantization point densities, for a given target entropy. In this work, quantizers are specified in terms of their point densities, i.e., the reciprocal value of the step-size of the quantizer.

II. Spherical Trellis-Coded Quantization

We start this section by introducing the mathematical problem statement of quantization in parametric audio coding based on a perceptual relevant distortion measure. Let the audio signal x be represented at sample time n for a given segment by $x(n) \approx \sum_{l=1}^{L} a_l \sin(\nu_l n + \phi_l)$, where L is the number of sinusoidal components and a_l, ϕ_l, ν_l are amplitude, phase and frequency of the l'th component, respectively. Here $a_l \geq 0$ 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 N, or statistical independence between the components, the total expected distortion can be approximated as the sum over the L expected distortions for the individual

M. H. Larsen, M. G. Christensen and S. H. Jensen are with the Department of Electronic Systems, Aalborg University, Denmark. (phone: +45 9635 8684, fax: +45 9635 1583 and email: {mhl,mgc,shj}@es.aau.dk)

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

components denoted $E[D]$, with $E[\cdot]$ being the expectation operator. Therefore, we will in the rest of this letter be concerned with the quantization of a single set of parameters (a, ϕ, ν) and drop the subscript l. A perceptual distortion that has proven successful for audio coding applications (see, e.g., [1], [4], [5], [11], [12]) was introduced in [13], and is defined as

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

where $E(\omega)$ denotes the discrete-time Fourier transform of the windowed error, i.e.,

$$E(\omega) = \sum_{n=n_0}^{n_0+N-1} w(n)(x(n) - \tilde{x}(n))e^{-j\omega n}, \tag{2}$$

where $w(n)$ is the window, $\mu_{x(a,\phi,\nu)}(\omega)$ is the perceptual weighting function that is calculated from the audio signal segment x that is parametrized by (a, ϕ, ν) and $\tilde{x}$ is the reconstructed audio signal based on the quantized parameters $(\tilde{a}, \tilde{\phi}, \tilde{\nu})$. The aim of this work is to construct quantizers for (a, ϕ, ν) subject to an entropy constraint such that the perceptual distortion in (1) is minimized. For an overview of methods for modeling an audio segments using sinusoids such that (1) is minimized, we refer the interested reader to [14] and the references therein. Assuming a large N, high-resolution and smooth masking curve, the perceptual distortion can be approximated as (see e.g. [4])

$$\begin{aligned} D \approx\ & \frac{\mu_{x(a,\phi,\nu)}}{2}\Bigg(\|w\|^2(a^2+\tilde{a}^2) - 2a\tilde{a} \\ & \times \sum_{n=n_0}^{n_0+N-1} w^2(n)\cos\Big((\nu-\tilde{\nu})n + \phi - \tilde{\phi}\Big)\Bigg), \end{aligned} \tag{3}$$

where $\|w\|^2 = \sum_{n=n_0}^{n_0+N-1} w^2(n)$. Here, we have assumed that N is large such that the perceptual weighting function reduces to a scaling $\mu_{x(a,\phi,\nu)}$ that depends on the realization $x(a, \phi, \nu)$. Similarly to [2]–[4], we in this work 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 difficult problem.

Assuming that the phase is defined in the middle of the segment, $n_0 = -\frac{N}{2}$, and applying a truncated Taylor expansion around zero, $\cos(y) \approx 1 - y^2/2$, the distortion (3) can be written as

$$D \approx \frac{\mu_{x(a,\phi,\nu)}}{2\|w\|^{-2}}\left((a-\tilde{a})^2 + a\tilde{a}\left((\phi-\tilde{\phi})^2 + (\nu-\tilde{\nu})^2\sigma^2\right)\right), \tag{4}$$

with $\sigma^2 = \frac{1}{\|w\|^2}\sum_{n=-N/2}^{N/2-1} w^2(n)n^2$. We observe from (4), that we can quantize the amplitude, phase and frequency individually using the l_2-norm, when assuming high-resolution, $a\tilde{a} \approx a^2$. This provides inspiration to the proposed STCQ coding scheme consisting of three TCQs, for a, ϕ and ν respectively. In deriving the optimal STCQ design, we need an expression for the expected distortion for the TCQ. However, an exact expression for the expected distortion does not exist for the TCQ. Therefore, we will employ the approximation proposed in [10] where the expected distortion is written as the distortion for a uniform quantizer corrected by a factor Γ. This factor depends on the trellis structure, dimension and the number of states and is roughly independent of the source pdf and encoding rate. Additionally, assuming high-resolution the distortion for one Voronoi region in a TCQ can be written as

$$\int_{\tilde{y}-\frac{g^{-1}}{2}}^{\tilde{y}+\frac{g^{-1}}{2}} \Gamma(y-\tilde{y})^2 dy = \Gamma\frac{g_y^{-3}}{12}, \tag{5}$$

where g_y is the quantization point density and y and $\tilde{y}$ are the source realization and its reconstruction, respectively. In Table I the factor Γ has numerically been determined by simulations. From (4) and the expected distortion for

TABLE I
Γ-VALUES FOR VARIOUS NUMBER OF DIMENSIONS AND STATES DETERMINED BY SIMULATION.

		Number of states			
		4	16	256	512
Dim.	10^1	3.5102	3.5976	3.8780	3.9314
	10^2	3.2107	3.1347	3.1019	3.1091
	10^3	3.1872	3.0714	2.9908	2.9752
	10^4	3.1829	3.0633	2.9745	2.9629

each of the three TCQs, as written in (5), we can now determine the expected distortion for one Voronoi region c, i.e.,

$$\begin{aligned} \iiint_c D d\nu d\phi da &= \frac{\mu_{x(a,\phi,\nu)}}{2\|w\|^{-2}}\int_{\tilde{a}-\frac{g_a^{-1}}{2}}^{\tilde{a}+\frac{g_a^{-1}}{2}}\int_{\tilde{\phi}-\frac{g_\phi^{-1}}{2}}^{\tilde{\phi}+\frac{g_\phi^{-1}}{2}}\int_{\tilde{\nu}-\frac{g_\nu^{-1}}{2}}^{\tilde{\nu}+\frac{g_\nu^{-1}}{2}} \\ &\times\left(\Gamma_a(a-\tilde{a})^2 + a\tilde{a}\left(\Gamma_\phi(\phi-\tilde{\phi})^2 + \Gamma_\nu(\nu-\tilde{\nu})^2\sigma^2\right)\right) da d\phi d\nu \\ &= g_a^{-1}g_\phi^{-1}g_\nu^{-1}\frac{\mu_{x(a,\phi,\nu)}}{24\|w\|^{-2}}\Gamma\left(g_a^{-2} + \tilde{a}^2\left(g_\phi^{-2} + \sigma^2 g_\nu^{-2}\right)\right), \end{aligned} \tag{6}$$

where we assume that $\mu_{x(a,\phi,\nu)}$ is constant over c and, for simplicity, we also assume an equal trellis structure, $\Gamma = \Gamma_{\{a,\phi,\nu\}}$. Here, the quantization point densities for a, ϕ, ν are written as $g_{\{a,\phi,\nu\}}$, though it should be noted that they at this point still depend on a, ϕ, ν but for now, we shall omit this in our notation. The expected distortion can then be written as,

$$E[D] = \sum_{i_a\in I_a}\sum_{i_\phi\in I_\phi}\sum_{i_\nu\in I_\nu}\iiint_c f(a,\phi,\nu) D d\nu d\phi da \tag{7}$$

$$\approx \sum_{i_a\in I_a}\sum_{i_\phi\in I_\phi}\sum_{i_\nu\in I_\nu}\frac{\Pr(\tilde{a}_{i_a}, \tilde{\phi}_{i_\phi}, \tilde{\nu}_{i_\nu})}{g_a^{-1}g_\phi^{-1}g_\nu^{-1}}\iiint_c D d\nu d\phi da \tag{8}$$

$$\begin{aligned} \approx & \frac{\|w\|^2}{24}\Gamma\iiint f(a,\phi,\nu)\mu_{x(a,\phi,\nu)}\big(g_a^{-2} \\ & + \tilde{a}^2(g_\phi^{-2} + \sigma^2 g_\nu^{-2})\big) d\nu d\phi da, \end{aligned} \tag{9}$$

where $f(a, \phi, \nu)$ is the source pdf and $\Pr(\tilde{a}_{i_a}, \tilde{\phi}_{i_\phi}, \tilde{\nu}_{i_\nu})$ is the probability mass function of the reconstruction points. Here we assumed high-resolution such that

$\Pr(\tilde{a}_{i_a}, \tilde{\phi}_{i_\phi}, \tilde{\nu}_{i_\nu}) = f(\tilde{a}_{i_a}, \tilde{\phi}_{i_\phi}, \tilde{\nu}_{i_\nu}) g_a^{-1} g_\phi^{-1} g_\nu^{-1}$. The entropy of a single TCQ is given by $H \approx h(Y) - 1 + \int f_Y(y) \log_2(g_y) dy$, where $h(Y)$ is the differential entropy of the source Y and $f_Y(y)$ is its pdf. Note that we here denote random variables by upper case letters whereas realizations are denoted by lower case. Assuming that the amplitude, phase and frequency are independent, we can then write the entropy as

$$H \approx h(A, \Phi, \Upsilon) - 3 + \iiint f(a, \phi, \nu) \log(g_a g_\phi g_\nu) da d\phi d\nu. \tag{10}$$

To simplify the notation, we get rid of the differential entropy of the source, $h(A, \Phi, \Upsilon)$, by introducing $\tilde{H} = H - h(A, \Phi, \Upsilon)$ and we write the cost function $J = E[D] + \lambda \tilde{H}$, where λ is the Lagrange multiplier. We minimize the cost function by taking the derivative with respect to g_a, g_ϕ and g_ν and set to zero, whereby we obtain

$$g_a = \left(\frac{\Gamma \mu_{x(a,\phi,\nu)} \|w\|^2}{12 \lambda \log_2(e)} \right)^{1/2} \tag{11}$$

$$g_\phi = \left(\frac{\Gamma \mu_{x(a,\phi,\nu)} \tilde{a}^2 \|w\|^2}{12 \lambda \log_2(e)} \right)^{1/2} \tag{12}$$

$$g_\nu = \left(\frac{\Gamma \mu_{x(a,\phi,\nu)} \tilde{a}^2 \sigma^2 \|w\|^2}{12 \lambda \log_2(e)} \right)^{1/2}. \tag{13}$$

Inserting the three last equations into the definition of $\tilde{H}$, we can express the optimal λ^*, as

$$\lambda^* = \frac{\Gamma \|w\|^2 2^{\psi} 2^{\frac{2}{3}\left(2\varrho + \log_2(\sigma) - 3 - \tilde{H}\right)}}{12 \log_2(e)}, \tag{14}$$

where $\varrho = \int f_A(a) \log_2(a) da$ with $f_A(a)$ being the pdf of A and $\psi = \iiint f(a, \phi, \nu) \log_2(\mu_{x(a,\phi,\nu)}) da d\phi d\nu$. Substituting λ^* into the expression of g_a, g_ϕ and g_ν, we get

$$g_a = \left(\frac{\mu_{x(a,\phi,\nu)}}{2^{\psi}} \right)^{\frac{1}{2}} 2^{\frac{1}{3}(\tilde{H} + 3 - \log_2(\sigma) - 2\varrho)} \tag{15}$$

$$g_\phi(\tilde{a}) = \tilde{a} g_a \tag{16}$$

$$g_\nu(\tilde{a}) = \sigma \tilde{a} g_a, \tag{17}$$

where the quantization point densities of ϕ and ν are now written as functions of $\tilde{a}$ to stress the dependencies on the quantized amplitude. Inserting (15)-(17) into (9) we can finally determine the expected distortion as

$$E[D] \approx \frac{\|w\|^2 \Gamma 2^{\frac{-2}{3}(\tilde{H} + 3 - 2\varrho - \frac{3}{2}\psi - \log(\sigma))}}{8}. \tag{18}$$

From (18) and the expected distortion in [4], we can calculate the theoretical performance gain between STCQ and the SQ as $10 \log_{10}(4/\Gamma)$ in dB. Using the values of Γ's from Table I, the distortion gain has been tabulated in Table II.

For the practical STCQ scheme, we observe from (15)-(17), that we must first encode a as $\tilde{a}$ before we can determine $g_\phi(\tilde{a})$ and $g_\nu(\tilde{a})$ and subsequently encode ϕ and ν. However, we see from equations (16) and (17) that $\tilde{a}$ simply scales the quantization point densities. In comparing the performance of various quantizers, it is also important to take into account the computational complexity. The encoding complexity per sinusoid of the STCQ is 12 uniform quantizers plus $3(2S+4)$ additions, 12 multiplications and $3S$ comparison of two real value numbers, where S is the number of states in the trellis. This in contrast to the SQ having a complexity of 3 uniform quantizers per sinusoid. The decoding complexity is very close to SQ decoding and thus negligible[1]. For a detailed discussion of the complexity and implementation details of the TCQ, we refer the interested reader to [9], [10].

TABLE II
THEORETICAL PERFORMANCE GAIN BETWEEN THE STCQ AND SQ [4] IN dB.

		Number of states			
		4	16	256	512
Dim.	10^1	0.5672	0.4605	0.1345	0.0752
	10^2	0.9546	1.0586	1.1044	1.0942
	10^3	0.9865	1.1472	1.2627	1.2855
	10^4	0.9923	1.1587	1.2864	1.3035

III. Experimental Results

The two following experiments are based on synthetic audio, generated based on a statistical model similar to that employed in [4]. 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 $[0, 2\pi)$. Note that the amplitudes, phase and frequency are independent. In this work we focus on the performance gain that can be achieved by joint quantization and we will therefore, for simplicity, set the perceptual weighting function to one and use a rectangular window with $N = 1023$. We remark that the trellis is initialized in zero-state and that no side information is needed for the trellis. In a first experiment, we illustrate the impact of the number of sinusoids that are jointly quantized and the number of states in the trellis on the performance of the quantizers. We generate 100,000 triplets (a,ϕ,ν), set the target entropy to 20 bit/sinusoid and quantize the triplets according to (15)-(17). The performance of the STCQ for four different number of states are shown in Fig. 1 together with the corresponding SQ, all evaluated by quantizing the triplets and then calculating the average distortion using (3). From Fig. 1 it can be seen, that we gain about 1 dB for low dimensions and 1.7 dB at high dimensions compared to the SQ of [4]. Furthermore, it can be seen that it is not beneficial to have a high number of states at low dimensions, but this may be improved by, e.g., a tail biting trellis. In the second experiment, we compare the performance of the STCQ to the SQ and the theoretical

[1]Another coding method is to construct a tensor product of the amplitude-, phase- and frequency trellises and use the STCQ encoding procedure, except that the Viterbi alg. is now applied on the product trellis. We have investigated the performance difference between this methods and the STCQ, and concluded that only at low bit-rate (between 5 and 10 bits/sinusoid) this method is slightly better the STCQ. Taking the complexity into an account, STCQ is favoured.

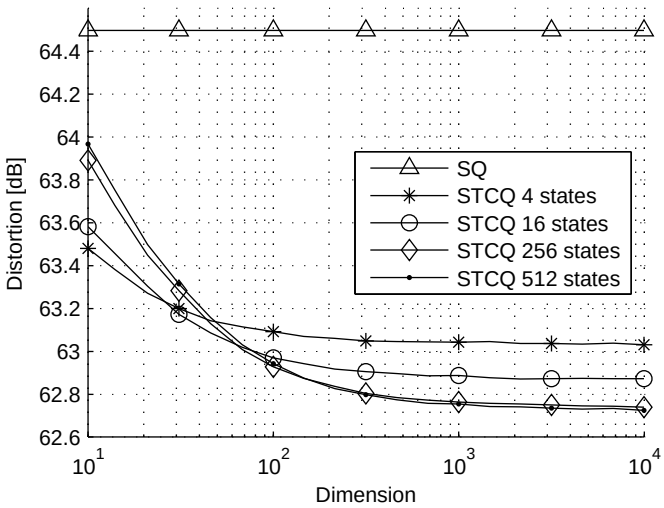

Fig. 1. The performance of the SQ and STCQ for four different number of states.

expected distortion for both STCQ and SQ. Again, we generate 100,000 triplets and jointly quantize 1000 sinusoids parameters using a 256 state STCQ. The distortion is calculated as in the previous experiment. The comparison for several target entropies are shown on Fig. 2. From this figure, it can be seen that STCQ outperforms the SQ and that the performance of the STCQ converges towards the theoretical performance. This verifies that (18) is a valid approximation. A likely explanation for the constant performance of the STCQ and the SQ in the interval 0 to 5 bit is that the distortion is plotted as a function of the target entropy. The same phenomenon was observed in [4]. Simulations indicate that, when plotted as a function of the resulting entropy, the distortion actually decreases for low entropies.

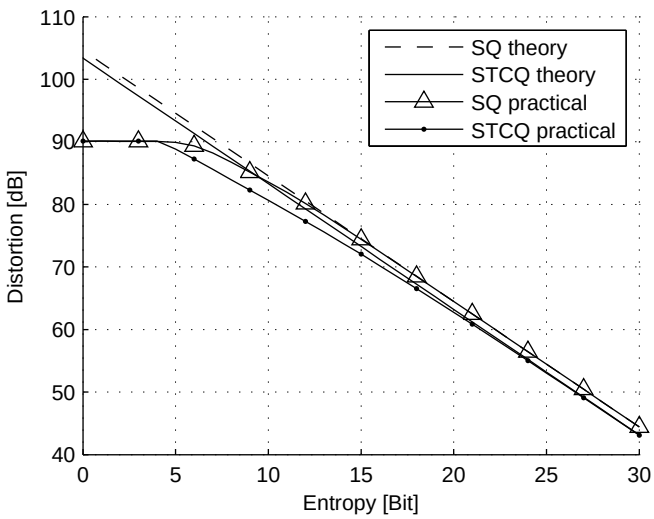

Fig. 2. The rate distortion performance of the STCQ (solid dotted line) and the SQ (solid line with triangles) together with the theoretical expected performance of both the STCQ (solid line) and the SQ (dashed line) [4].

IV. Conclusions

We have proposed spherical trellis-coded quantization of sinusoids that is suitable for variable rate parametric audio coding, where the number of sinusoids is variable. A prominent feature of the proposed scheme is that the sinusoidal parameters are quantized jointly in a computationally efficient manner. Furthermore, we have derived analytical expressions for the optimal design and the expected perceptual distortion under high-resolution assumptions. Experiments have shown a significant performance gain compared to the spherical quantization scheme of [4]. This performance gain comes at the cost of a slight increase in encoding complexity. Furthermore, the impact on the performance when increasing the number of states in the trellis and the number of sinusoids that are jointly quantized has been investigated in experiments, and it has been found that the performance of the proposed scheme improves as the number of states and dimensions is increased.

References

[1] R. Heusdens and J. Jensen, "Jointly optimal time segmentation, component selection and quantization for sinusoidal coding of audio and speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, March 2005, pp. 193–196.

[2] R. Vafin and W. Kleijn, "Entropy-constrained polar quantization and its application to audio coding," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 2, pp. 220–232, March 2005.

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

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

[5] R. Heusdens, J. Jensen, W. B. Kleijn, V. Kot, O. A. Niamut, S. van de Par, N. H. van Schijndel, and R. Vafin, "Bit-rate scalable intra-frame sinusoidal audio coding based on rate-distortion optimisation," *J. Audio Eng. Soc.*, vol. 54, no. 3, pp. 167–188, March 2006.

[6] P. Lupini and V. Cuperman, "Nonsquare transform vector quantization," *IEEE Signal Processing Lett.*, vol. 3, no. 1, pp. 1–3, Jan. 1996.

[7] J. Lindblom and P. Hedelin, "Variable-dimension quantization of sinusoidal amplitudes using Gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, May 2004, pp. 153–156.

[8] M. Marcellin, "On entropy-constrained trellis coded quantization," *IEEE Trans. Commun.*, vol. 42, no. 1, pp. 14–16, Jan. 1994.

[9] M. W. Marcellin, "Trellis coded quantization: An efficient technique for data compression," Ph.D. dissertation, College of Taxas A&M University, 1987.

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

[11] M. G. Christensen and S. van de Par, "Efficient parametric coding of transients," *IEEE Trans. Audio, Speech and Lang. Processing*, vol. 14, no. 4, pp. 1340–1351, Jul. 2006.

[12] M. G. Christensen, A. Jakobsson, S. V. Andersen, and S. H. Jensen, "Amplitude modulated sinusoidal signal decomposition for audio," *IEEE Signal Processing Lett.*, vol. 13, no. 7, pp. 389–392, Jul. 2006.

[13] 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," *EURASIP J. on Applied Signal Processing*, vol. 9, pp. 1292–1304, June 2005.

[14] 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, Jan. 2006.