Robust Parametric Audio Coding Using Multiple Description Coding

Jensen, Jesper Rindom; Christensen, Mads Græsbøll; Larsen, Morten Holm; Jensen, Søren Holdt; Larsen, Torben

Published in:
IEEE Signal Processing Letters

DOI (link to publication from Publisher):
10.1109/LSP.2009.2030852

Publication date:
2009

Document Version
Accepted author manuscript, peer reviewed version

Link to publication from Aalborg University

Citation for published version (APA):
Robust Parametric Audio Coding Using Multiple Description Coding

Jesper Rindom Jensen, Mads Græsbøll Christensen, Member, IEEE, Morten Holm Jensen, Member, IEEE, Søren Holdt Jensen, Senior Member, IEEE, Torben Larsen, Member, IEEE

Abstract—We propose a new multiple description spherical quantization with repetitively coded amplitudes (MDSQRA) scheme suited for quantization of sinusoidal parameters. The quantization scheme is constituted by a set of spherical quantizers inspired by the multiple description spherical trellis-coded quantization (MDSTCQ) scheme. In this scheme, we apply repetitive coding on the amplitudes, while multiple description coding are applied on the phases and frequencies. Thereby, MDSQRA becomes directly implementable, as opposed to MDSTCQ, since the phase and frequency quantizers depend on the amplitudes which have dissimilar descriptions in MDSTCQ. Furthermore, we implement MDSQRA into a perceptual matching pursuit based sinusoidal audio coder. Finally, we evaluate MDSQRA through perceptual distortion measurements and MUSHRA listening tests. The tests show that MDSQRA outperforms MDSTCQ with respect to a expected perceptual distortion measure. The same results are obtained through the MUSHRA tests performed on sound clips coded using MDSQRA and MDSTCQ.

Index Terms—Perceptual audio coding, multiple description coding, pre- and post-filtering, sinusoidal parametric coding, spherical quantization

I. INTRODUCTION

The reduction of the bit rate for a given audio quality has been subject to comprehensive research in the past few decades. It is desired that the bit rate is reduced without compromising the audio quality meaning it is not sufficient to just remove the statistical redundancies of the audio signals, if a large compression factor is desired. A remedy for attaining higher compression ratios is to use perceptual audio coding. These audio coders take into account that certain parts of audio signals are inaudible to the human ear. One way of discarding such irrelevancies is to determine and apply a masking threshold below which signal components are inaudible. The masking threshold is dependent on the time, frequency and amplitude characteristics of the audio signal [1]. There exist three common classes of perceptual audio coding schemes, namely, sub-band coding, transform coding and parametric coding.

Parametric coding exploits, that many audio signals can be represented using a few perceptually important parameters. A well known parametric coding scheme is sinusoidal coding where the audio signal is described as sums of sinusoids each being characterized by an amplitude, a phase and a frequency. Sinusoidal coding entails the task of estimating the parameters best describing the audio signal and quantization of these relevant parameters. There exist several computationally efficient approaches for estimation of the sinusoidal parameters that minimize a perceptual distortion measure. For an overview of such methods see e.g. [2].

In this paper, we propose a novel parametric audio coding framework using our proposed multiple description spherical quantization with repetitively coded amplitudes (MDSQRA).

Copyright © 2008 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permission@ieee.org. J.R. Jensen, M.G. Christensen, S.H. Jensen and T. Larsen are with the Department of Electronic Systems, Aalborg University, Denmark, (phone: +45 9940 8617, fax: +45 9635 1583 and email: {jesperj,mgc,shj,tl}@es.au.dk). M.H. Jensen is with Widex A/S, Denmark. (email: holm@ ieee.org) M.G. Christensen was supported by the Parametric Audio Processing project, Danish Research Council for Technology and Production Sciences grant no. 274–06–0521.

S.H. Jensen was partly supported by the Danish Technical Research Council, through the framework project Intelligent Sound, www.intelligentsound.org (STVF No. 26-04-0092)
parametric audio coding in the form of sinusoidal coding and we propose the MDSQRA quantization scheme. The proposed quantization scheme is implemented in a parametric audio coder supported by experimental results in Section 3, followed by a conclusion in Section 4.

II. PROPOSED METHOD

We use the following sinusoidal model for the sinusoidal coding. For a model order \( L \) and a time segment \( n = 1, \ldots, N \) of an audio signal \( x[n] \) the model is given by

\[
\hat{x}[n] = \sum_{l=1}^{L} a_l \cos(\nu_l n + \phi_l)
\]

(1)

where \( a_l, \nu_l \) and \( \phi_l \) are the amplitude, frequency and phase characterizing the \( l \)th sinusoid, respectively. Typically, the time segments are windowed and overlapping. For each time segment the sinusoidal parameters have to be estimated. As an example this can be done using the perceptual matching pursuit (PMP) algorithm [11] (see also [2]).

After the estimation of the sinusoidal parameters they should be quantized before transmission. We therefore introduce MDSQRA of the sinusoidal parameters. The quantizers are derived on the basis of minimization of a perceptual distortion measure given an entropy constraint. The total expected distortion consists of the sum of the distortions from quantization of the individual sinusoidal components and their cross-terms. In this paper, however, we assume sufficiently long windows, such that the individual sinusoids become orthogonal which allows us to neglect the cross-term distortion. Therefore, we only consider quantization of one set of parameters.

The perceptual distortion measure used in these derivations is based on a perceptual weighting function \( \mu(x, \phi, \nu) \) and it is given and can be approximated by (see, e.g., [12])

\[
D = \frac{1}{2\pi} \int_{0}^{2\pi} \mu_{x, \nu}(\omega) |\epsilon(\omega)|^2d\omega
\]

\[
\approx \frac{\mu_{x, \nu, \phi}(\omega)}{2|w|^2} ((a - \tilde{a})^2 + a\tilde{a}((\phi - \tilde{\phi})^2 + (\nu - \tilde{\nu})^2\sigma^2))
\]

(3)

where \( \sigma^2 = \frac{1}{|w|^2} \sum_{n=0}^{N-1} w^2(n)n^2 \) is the set of quantized sinusoidal parameters, and (3) follows from (4). The DTFT of the windowed error is defined as

\[
\epsilon(\omega) = \sum_{n=0}^{N-1} w(n)(x(n) - \hat{x}(n))e^{-j\omega n}
\]

(4)

The multiple description quantizers considered here generate two descriptions. If only one description is received in the decoder, a low quality description is obtained with a side distortion \( D_s \) where \( s = \{1, 2\} \). If both descriptions are received the distortion reduces to the central distortion \( D_0 \). In this work, the quantizers are designed to make a sinusoidal audio coder robust against transmission over a packet erasure channel where packets are lost independently with probability \( p \). The expected distortion for such a channel is given by

\[
E[D] = (1 - p)^2E[D_0] + 2p(1 - p)E[D_s] + p^2\sigma^2
\]

(5)

where \( \sigma^2 \) is the variance of the audio signal and the side distortion is balanced (i.e. \( E[D_1] = E[D_2] = E[D_0] \)).

The distortions in (5) are found as the sum of the distortions from the amplitude, frequency and phase quantizers. For the amplitude quantizer we know that \( E[D_a] = E[D_s] \) since it uses repetition coding, whereas \( E[D_0] \approx \frac{E[D_0]}{2N^2} \) for the frequency and phase quantizers since they are based on the modified multiple description scalar quantization scheme in [13]. Knowing this, the following expressions for the expected distortions can be obtained from (3) as

\[
E[D_a] = \frac{|w|^2}{24} \int \int \int f_{A, \nu, \phi}(a, \nu, \phi) \mu_{x, \nu, \phi}(a, \nu, \phi, \nu) (g_a^2 + \tilde{a}^2 g_a^{-2} + \sigma^2 g_a^{-2}) d\nu d\phi
\]

\[
E[D_0] = \frac{|w|^2}{96} \int \int \int f_{A, \nu, \phi}(a, \nu, \phi) \mu_{x, \nu, \phi}(a, \nu, \phi, \nu) (4g_a^2 + \tilde{a}^2 \left( \frac{g_o^2}{N_o^2} + \sigma^2 \frac{g_o^{-2}}{N_o^2} \right)) d\nu d\phi
\]

(6)

(7)

Following, the expected distortions in (6) and (7) are inserted into (5) which is then minimized subject to an entropy constraint using the Lagrange multiplier method (see [14] for details). This gives the following expressions for the squared optimal quantization point densities and the optimal number of refined reconstruction points

\[
g_a^2 = \left( 1 + \frac{p}{4\rho} \right)^{-\frac{1}{2}} \frac{\mu_{x, \nu, \phi}(a, \nu, \phi)}{N^2} \left( \tilde{a} - \log_2 \left( \frac{\rho}{2} \right) - 2b(a) \right)
\]

\[
g_o^2 = \frac{4p}{1 + p} g_a^2 \text{ and } g_o^2 = \tilde{a}^2 \sigma^2 \frac{4p}{1 + p} g_a^2
\]

\[
N = \left( 1 - \frac{p}{8\rho} \right)^{-\frac{1}{2}}
\]

(8)

(9)

(10)

where \( \tilde{H}_s = H_s - h(A, \Phi, \Psi) \), \( H_s \) is the entropy of one description, \( h(A, \Phi, \Psi) \) is the joint differential entropy, \( N \) is the number of refined reconstruction points, and

\[
b(a) = \int f_{A}(a) \log_2(\rho(a, \nu, \phi) - 2b(a)) da
\]

\[
\rho(a, \nu, \phi) = \int \int f_{A, \nu, \phi}(a, \nu, \phi) \log_2(\mu_{x, \nu, \phi}(a, \nu, \phi)) d\nu d\phi
\]

(11)

(12)

with \( f_{A}(a) \) being the probability density function of the amplitude and \( f_{A, \nu, \phi}(a, \nu, \phi) \) being the joint probability density function of the amplitude, phase, and frequency. The resulting
Hann windowed segments were used in the sinusoidal coder phases and frequencies for each sound clip. Overlapping von $\rho$ probability in Fig. 1. The chosen values for $\phi, \nu$, and $\nu$ are 3 and 4, which is assumed to be realistic values according to [5]. As it can be seen from the plot, the expected distortion can be found for MDSTCQ which is also shown in Tab. I. Likewise, as shown in [7], the expected distortion can be found for MDSTCQ which is also shown in Tab. I.

III. EXPERIMENTAL RESULTS

This section describes the evaluation of the proposed robust parametric audio coder. First, MDSQRA and MDSTCQ are compared with respect to their expected perceptual distortion. For this purpose we use the theoretical expressions presented in Tab. I. The factors that differ between the expected distortions for the two quantizers are

$$k_{\text{MDSQRA}} = \frac{3}{24} (1 - p^2) N^\frac{1}{2} \left( \frac{4p}{1 + p} \right)^{\frac{1}{2}}$$

$$k_{\text{MDSTCQ}} = \sqrt{\frac{5p}{8}} (1 - p)^\frac{1}{2}$$

The two factors are plotted as functions of the packet loss probability in Fig. 1. The chosen values for $\Gamma$ is 3 and 4, respectively, which is assumed to be realistic values according to [5]. As it can be seen from the plot, the expected distortion for MDSQRA is lower than MDSTCQ for the whole range of packet loss probabilities.

Furthermore, we have conducted two MUSHRA listening [15] tests on the MDSQRA-based parametric audio coder with 13 non-expert listeners. In the first MUSHRA test we investigate the performance gain obtained when receiving two descriptions compared to when only one description is received. The quantizers are designed at a packet loss probability $p = 10\%$ which gives $N = 3$. The unknown parameters in the quantizer designs, i.e. the differential entropies, $b(a)$ and $\rho(a, \phi, \nu)$, where estimated from the unquantized amplitudes, phases and frequencies for each sound clip. Overlapping von Hann windowed segments were used in the sinusoidal coder with a window length of 30 ms. Each segment was modeled by $L = 30$ sinusoids and coded at a total bit-rate of 48 kbit/s including both descriptions and the perceptual weights. We used eight sound clips coded using this setup in the listening tests. Four of the clips contained instrumental music from, a bass, a piano, a violin and a xylophone whereas the other four contained music from Abba, Eric Clapton, Tracy Chapman and a soprano. It should be stressed that these sounds clips are complicated and not easily modeled with only sinusoids. However, the applied audio coder only contains a sinusoidal coder for conceptual simplicity whereas transient and noise components are not modeled. We remark that the considered coding scheme could be extended to account for this by, for example, using the coder in [10] as a waveform approximating residual coder as proposed in [16].

In the first test we quantized the sound clips using MDSQRA. The sound clips were reconstructed in two versions; one where we always receive both descriptions and one where we always only receive one description. Besides that, the listeners were presented with a 3.5 kHz low pass filtered anchor signal and a hidden reference signal. The test results are presented in Fig. 2 where the score 100 corresponds to "Imperceptible" and the score 0 corresponds to "Very annoying" according to the ITU-R 5-grade impairment scale. It is clearly seen from the figure, that there is a coding gain when receiving two descriptions compared to just receiving one which confirms the desired effect of MDSQRA. Note, that the anchor signal scores higher than the coded versions. This is explained by the fact that most of the used sound clips consisted of mostly low frequency content.

In the second test we compared the MDSQRA scheme with two SQ schemes. In the first SQ scheme we coded the amplitudes, frequencies and phases using pure repetition

<table>
<thead>
<tr>
<th>Test</th>
<th>Expected distortion</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDSQRA</td>
<td>$\frac{3}{24} (1 - p^2) N^\frac{1}{2} \left( \frac{4p}{1 + p} \right)^{\frac{1}{2}}$</td>
</tr>
<tr>
<td>MDSTCQ</td>
<td>$\sqrt{\frac{5p}{8}} (1 - p)^\frac{1}{2}$</td>
</tr>
</tbody>
</table>

**TABLE I**

**EXPECTED DISTORTION FOR MDSQRA AND MDSTCQ.**

![Fig. 1. Comparison of the expected distortion scaling factors for MDSQRA and MDSTCQ.](image1.png)

![Fig. 2. MUSHRA test results on multiple description coding gain averaged over all eight sound clips. The four versions of the clips appear in the following order: Anchor, Hidden ref., 1 description, and 2 descriptions.](image2.png)
coding and send two equal descriptions whereas in the other SQ scheme we send only one description. The coder setup was the same as in the first MUSHRA test. We coded the same eight sound clips using the mentioned setup and quantizers. For each quantization scheme we reconstructed the sound clips with packet loss probabilities of 5 % and 10 %, respectively. Besides the coded clips the listeners were presented to a 3.5 kHz low pass filtered anchor signal and a hidden reference signal. We have depicted the outcome of this test in Fig. 3. It can be seen from this test that there is a clear advantage of using multiple description coding compared to sending only one description when packet losses are present. However, the difference between MDSQRA and repetitive coded SQ is small. If we look at the difference in the scores between the two at $p = 10 \%$ we get that the mean of this is 3.3 and the 95 % confidence interval is located ±3.3 around this value. However, it is expected that the difference is small at increasing packet loss probabilities since the effect of the central description is degraded. Furthermore, the first listening test showed that there is performance gain when using MDSQRA compared to single description SQ at low packet-loss probabilities.

IV. Conclusion

In this paper, we considered a novel multiple description quantization scheme for quantization of sinusoidal parameters, namely multiple description spherical quantization with repetitively coded amplitudes and its application to real audio signals. The proposed quantizers are simple, closed form and computationally efficient in comparison with SQ. Compared to MDSTCQ as proposed in [7] MDSQRA does not introduce a suboptimality in the implementation. We compared the two quantization schemes and showed that MDSQRA performs better than MDSTCQ with respect to perceptual distortion. Furthermore, we implemented MDSQRA in a sinusoidal audio coder. As a proof of concept, we compared the performance of the MDSQRA-based audio coder with a SQ-based audio coder through MUSHRA listening tests. The tests showed the following: 1) MDSQRA has the desired ability to improve robustness against packet losses and 2) MDSQRA has complementary descriptions such that receiving two descriptions improves the quality of the reconstructed sound clip compared to when only one description is received. The most significant results are, that the MDSQRA has a better performance than MDSTCQ with respect to the perceptual distortion and the computational complexity is comparable to SQ. Furthermore, it has been demonstrated through listening tests, that the proposed MDSQRA can improve the performance with respect to perceptual distortion, when packet losses are present, compared to a SQ scheme for real audio signals. Future work should consider generalization to more than two descriptions.

**REFERENCES**


