VECTORIAL SPECTRAL QUANTIZATION FOR AUDIO CODING

Adriana Vasilache and Henri Toukomaa

Nokia Research Center, Visiokatu 1, 33720 Tampere, Finland

ABSTRACT

The paper introduces a new coding methodology of the spectral modified discrete cosine transform (MDCT) coefficients of an audio signal. A lattice quantizer is used for each spectral sub-band, having the dimension equal to the size of the respective sub-band. The information that needs to be encoded consists of lattice codevector indexes, side information relative to the number the bits on which the indexes are represented and the integer exponents of the sub-band scaling factors. The nature of the side information, together with the parameterization of the quantization resolution allows the use of the method for a large range of bitrates e.g. for 44.1kHz sampled mono files, from 128 kbits/s down to 16 kbits/s. Subjective listening tests show similar performance of the proposed method to the advanced audio coding (AAC) codec for high bitrates (128 kbits down to 64 kbits/s) and clearly better performance for lower bitrates.

1. INTRODUCTION

The performance improvement of the spectral coefficients quantization is a very important issue within the coding of audio signals because the encoded bitstream is mainly formed of the bits related to it.

The quantization of the spectral coefficients within the AAC is performed using scalar quantization followed by entropy coding of some scale factors and of the scaled spectral coefficients. The entropy coding is performed as differential encoding using eleven possible fixed Huffman trees for the spectral coefficients and one tree for the scale factors. The scales for each band are set within two consecutive loops that make the process rather slow and do not take into account the inter-band correlations affecting the entropy coding. More detailed explanation on the implementation can be found in [1].

A clear performance improvement, with a complexity pay-off, has been obtained with the method proposed in [2] where the optimization of the scale factors was done keeping track of the correlations between consecutive spectral coefficients.

An alternative approach to account for these correlations is the use of vector quantization instead of the scalar one. The clear advantages of the vector quantization over its scalar counterpart have brought it into the attention of researchers, but practical issues like the size of the codebook or the complexity of the encoding algorithms have prevented earlier feasible results. The successful attempts have been mainly on the lower bitrates [3], but could not be extended to the whole domain of bitrates. Recent work based on the use of 4 dimensional lattice quantization has been presented in [4], but its results have been concentrated on a reduced bitrate domain.

The present study describes a spectral vector quantization technique based on lattices that make use of the advantages brought by the higher dimensional quantization while having a very low encoding complexity. The quantizer dimensions are equal to the sub-band sizes and at the quantization process a perceptual measure is minimized. Besides the constraint imposed by the limited amount of bits, an additional constraint is added during the quantization process, which avoids the setting to zero of a whole sub-band, to counteract some of the drawbacks of the perceptual model. As also observed in [5], the setting to zero of an entire sub-band, even if the overall (average or maximum) noise to mask ratio is below 1, may engender perceptual artifacts because the masking effects have been calculated on the original signal. These problems can in principle be alleviated through the introduction of a loop to consider the perceptual characteristics of the quantized signal, but it would increase significantly the complexity of the encoding. Another approach has been considered in [5] where the minimization of the error is done in the loudness domain. Although good results have been obtained, the combination with the use of the MDCT is still problematic. The use of the MDCT has been combined with spectral companding in [6] for speech signals whose spectral coefficients are modeled through a mixture of Gaussians. Our proposed method uses instead a single compressor function, modeling the data with a generalized Gaussian and chooses to improve the spectral quantization within the setup of the AAC encoding [1], to take advantage of the use of the MDCT. It provides very good encoding performance over a large domain of bitrates.

The paper presents first algorithmic details of the proposed method: quantization structure and encoding variables. The method has been tested through subjective listening tests for bitrates from 128kbits/s down to 16kbits/s and their results for 44.1kHz sampling frequency files are presented in a dedicated section. Conclusion and further work prospectives form the subject of the last section of the paper.

2. QUANTIZATION OF MDCT SPECTRAL COEFFICIENTS

The global encoding framework is similar to the one used in AAC. Within the bit pool mechanism, at each frame a given number of bits is available for the quantization of the MDCT coefficients grouped in several sub-bands, according to the perceptual model. Roughly, only half of the coefficients are actually quantized, the coefficients corresponding to the higher frequencies being set to zero. The number of spectral coefficients, the number of sub-bands and their lengths depend upon the sampling frequency of the input audio signal. We particularize in the following for the 44.1kHz case.

The MDCT coefficients from each sub-band i, are multiplied with b^{-s_i} and the result is further encoded. The encoding consists of scaling further the sub-bands by their experimental standard deviation value σ_i, companding the scaled coefficients and quantizing in a rectangular truncation of the lattice Z_n. Before quantization, but after companding, an additional multiplication with a factor $M \times \sigma_i$ is performed on the companded data, in order to adapt the quantization resolution to the initial energy of the sub-bands. The information to be encoded consists of the scale factor exponents $\{s_i\}$, the lattice

codevectors indexes and a side information providing the number of bits on which each index is represented. We denote in the following $\{s_i\}$ by scales. To be able to cover a wide range of bitrates by adapting the quantization resolution to the available number of bits, the base b used for the calculation of the down-scale factors, and the multiplicative factor M depend on the overall bitrate set by the user. Therefore, a single method, with only several parameter adjustments is used for the entire bitrate domain.

The scales are integers from a finite domain and they are entropy coded, same as the information relative to the number of bits on which the lattice codevectors are represented. The overall bitrate per frame depends on the choice of the scale factor values for each sub-band. The minimization of the quantization distortion versus the bitrate within the available number of bits per frame is performed following the algorithm presented separately in section 2.2. Next, we give details about the compressor function, quantization and codevector indexing.

2.1. Companding, quantization, and indexing

Division with an experimental standard deviation value σ_i normalizes each down-scaled sub-band. Experimentally, the probability density function of the n-dimensional normalized data in sub-bands is close to a generalized Gaussian function with shape factor α depicted in Figure 1:

$$p(\mathbf{x}) = \Big[\frac{\eta(\alpha,\sigma)}{2\Gamma(1/\alpha+1)}\Big]^n \exp\Big[-\eta(\alpha,\sigma)^\alpha \sum_{i=1}^{n} |x_i|^\alpha\Big], \qquad (1)$$

where $\sigma^2 = 1$ is the variance, Γ is the Gamma function, and the normalization constant $\eta(\alpha,\sigma)$ is $\eta(\alpha,\sigma) = \frac{1}{\sigma}\sqrt{\frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)}}$.

This function is used to infer the cumulative density function that engenders the scalar compressor function:

$$f(y) = \int_{-\infty}^{y} p_i(\mathbf{x})\mathbf{dx} \qquad (2)$$

where p_i is the marginal probability density function of $p(\mathbf{x})$.

The companded data has almost a uniform distribution and can be efficiently quantized using a lattice quantizer. The high-rate theoretically optimal compressor function [7]

$$f(y) = c\sigma_i \int_{-\infty}^{y} p_i^{\frac{n}{n+2}}(\mathbf{x})\mathbf{dx} \qquad (3)$$

has also been tested, but perceptually, for the common sub-band shape factor, better results have been obtained with the cumulative distribution function as compressor function. For memory saving reasons a single companding function, corresponding to $\alpha = 0.5$ has been considered for all the sub-bands.

To increase the quantization resolution, the companded data is additionally multiplied by the standard deviation of the sub-band times a factor, M, equal to 3 for bitrates greater or equal to 48kbits/s, and equal to 2.1 for bitrates less than 48kbits/s. The choice of the factor M as well the choice of the base b for high and low bitrates are issued from an experimental compromise of compression efficiency and encoding time and they are presented in Table 2 along with other settings for the encoder. A rectangular truncation of the Z_n lattice, $\Lambda = \{\mathbf{x} \in \mathbb{Z}^n | N(\mathbf{x}) \leq K\}$, is used for quantization. K is the norm of the truncation. The norm corresponding to the rectangular truncation is the maximum absolute value: $N(\mathbf{x}) = \max_i(|x_i|), \mathbf{x} = (x_1, ..., x_n) \in Z_n$. The data that must be encoded for the each sub-band i consist of the scale s_i which is an integer between 0 and 42,

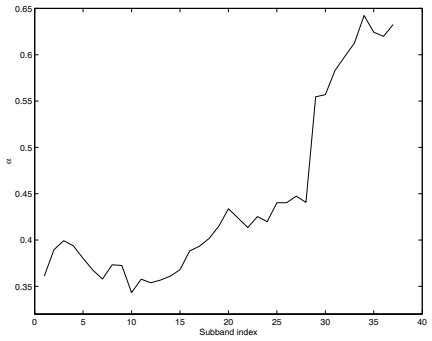

Fig. 1. Estimated shape factor value for sub-band data.

the lattice codevector index I_i and the norm n_i of the codevector, which is an integer between 0 and 141. The norms and the scale factors are entropy encoded. If the norm of a codevector is zero, the scale is not longer encoded, only the norm. The number of bits on which the indexes are represented can be calculated as:

$$Nbits = \lceil n \log_2(2 \cdot n_i + 1) \rceil \qquad (4)$$

if all the rectangular truncation is considered, or,

$$Nbits = \lceil \log_2(2 \cdot n_i + 1)^n - (2 \cdot n_i - 1)^n) \rceil \qquad (5)$$

if only the lattice vectors having norm n_i are considered. The variable n is the dimension of the quantization space, i.e. of the current sub-band and $\lceil \cdot \rceil$ is the closest integer to the argument, rounded toward infinity.

The indexing method for the codevector $\mathbf{x} = (x_1, \ldots, x_n)$ obtained in sub-band i, corresponding to Equation (4) has very low complexity, consisting only in the translation of the vector components by the norm n_i of the vector and constructing the index as:

$$I(\mathbf{x}) = \sum_{j=1}^{n} (x_j + n_i)(2n_i + 1)^{j-1}. \qquad (6)$$

The second indexing method is more complex and since it does not bring an important bitrate gain for the considered codevector sizes, the first method has been used.

From the point of view of memory requirements, the main storage load of the quantization method comes from the compressor function, which, in the present implementation, has higher x resolution in the central region and lower at the extremes, being tabulated and stored on 5kB.

2.2. Optimization algorithm

The goal of the optimization algorithm is to choose the scale value for each sub-band of the current frame such that the cumulated number of bits needed to encode the resulting codevector index, norm and scale for all the sub-bands is less than or equal to the available bitrate for the frame and the overall error ratio is as small as possible. The distortion measure is the ratio between the Euclidean distortion of quantization per sub-band over the allowed distortion for the considered sub-band. We do not use the average error ratio because, perceptually, the lowest average error ratio does not generally correspond to the best audio quality, as it has been experimentally observed. When a sub-band is set entirely to zero through the quantization process, even if the quantization distortion is much less than the

Sub-	Rate-distortion points					
band	R	D	R	D	R	D
SB1	$R_{1,1}$	$D_{1,1}$	$R_{1,...}$	$D_{1,...}$	R_{1,N_1}	D_{1,N_1}
SB2	$R_{2,1}$	$D_{2,1}$	$R_{2,...}$	$D_{2,...}$	R_{2,N_2}	D_{2,N_2}
...	...	...	...	...	...	...
SBn	$R_{n,1}$	$D_{n,1}$	$R_{n,...}$	$D_{n,...}$	R_{n,N_n}	D_{n,N_n}

Table 1. Ordered rate-distortion points.

Bitrate $\geq$ 48kbits/s	Bitrate $<$ 48kbits/s
$b = 1.45$	$b = 2.0$
M = 3	M = 2.1
Error ratio from definition	Error ratio from modified definition
Avoid zero	Do not avoid zeros

Table 2. General settings for the encoder.

Name	Description	Name	Description
es01	Vocal (S. Vega)	si01	Harpsichord
es02	German male speech	si02	Castanets
es03	English female speech	si03	Pitch pipe
sc01	Trumpet solo and orch.	sm01	Bagpipes
sc02	Classical orch. music	sm02	Glockenspiel
sc03	Contemp. pop music	sm03	Plucked strings

Table 3. Listening test samples.

allowed distortion, there might be audible artifacts in the resulting audio signal. This could be caused, for instance by the elimination of some of the maskers whose effect was actually calculated on the non-quantized signal. Therefore, we have chosen to give preference to a slightly larger average error ratio if it avoids the total setting to zero of a sub-band. This procedure is especially used for higher bitrates (higher than 48kbits/s).

The search space for the scale of a sub-band of dimension n is initialized with $\lfloor \log_b \sqrt{aD/n} \rfloor - 3$, where aD is the allowed quantization distortion per sub-band, issued from the perceptual model, and $\lfloor \cdot \rfloor$ represents the integer part, or the closest smaller integer to the argument. The optimization algorithm goes as follows. For each sub-band at most 20 rate-distortion points are evaluated, corresponding to the 19 scale values larger than the initial one plus the initial one. If there are not 20 scale values larger than the initial value, then only those available are considered. One rate-distortion point corresponds to the error ratio for the considered sub-band and the number of bits needed to encode the scale factor, lattice codevector index and norm (see Table 1, where $R_{i,j}$, $R_{i,j} < R_{i,j+1}$ are the number of bits and $D_{i,j}$, $D_{i,j} > D_{i,j+1}$, are the error ratios for $1 \leq i \leq n$ and $1 \leq j \leq N_i$). The number of bits for the entropy coded variables is estimated using a Shannon code. For each sub-band the rate-distortion points are sorted such that the bitrate is increasing and the error ratio decreases. In case this rule is violated, if for higher bitrate the error ratio does not decrease, the point with the higher bitrate is eliminated. When calculating the error ratio there are two possible approaches: one calculates the real error ratio from its definition, and the second one sets to zero the error ratio if the allowed distortion is larger than the energy of the signal in the considered sub-band. We denote the first approach by "definition" and the second one by "modified definition". The algorithm is initialized with the rate-distortion points corresponding to the lowest error ratios (equivalent to the highest number of bits). The first approach has been chosen both for the high and low bitrate cases. If the number of bits available for the frame the algorithm is less than the sum of the number of bits corresponding to the initial set of rate-distortion points, the algorithm decreases the number of bits in a greedy manner by reducing the number of bits for some of the sub-bands. The eligible sub-bands for number of bits reduction are those using more than 1 bit (1 bit corresponding to all the coefficients in that sub-band being set to zero). For overall bitrates larger or equal to 48kbits/s an additional constraint of having the number of bits larger than 1 is imposed in every sub-band. For each eligible sub-band, the gradient corresponding to the advancement of one pair to the left is calculated, and the one having maximum decrease in bitrate with lowest increase in distortion is selected. Then, the resulting total bitrate is checked, and so on.

We summarize the conditions and constant settings for low and high bitrates in Table 2.

3. LISTENING TESTS RESULTS

The proposed method was compared against the quantization procedure from the MPEG4-AAC codec, in a MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) [8]. A particularity of the AAC codec framework was the 11kHz bandwidth considered for quantization for all the bitrates. Three experiments were designed: for high bitrates (128kbits/s, 96kbits/s, 64kbits/s), for moderate bitrates (64kbits/s, 48kbits/s, 32kbits/s), for low bitrates (32kbits/s, 24kbits/s, 16kbits/s). There was also a training experiment for the listeners, covering all the audio quality range. The files used in the tests are listed in Table 3. The files es01 and sm01 were used only in the training experiment and the remaining files were used in each of the three testing experiments. The mono files are sampled at 44.1kHz. The number of expert listeners was 10 (high bitrates), 11 (moderate bitrates) and 9 (low bitrates).

Table 4 presents method comparison results issued through the Student T-test from the listening tests. "Cmpnd" stands for the proposed method using companding, AAC for the AAC coder and the numbers attached to them represent the bitrate in kbits/s. The grade '1' stands for 'first method is statistically better', '0' stands for statistically equal performance and '-1' for 'first method is statistically worse'. For high bitrates (128, 96, 64 kbits/s) the proposed method gives similar audio quality to the AAC. For bitrates of lower or equal to 48kbits/s the proposed method achieves better performance compared to the quantization method from the original AAC codec. Furthermore, for low bitrates, there is an important quality improvement, equivalent to bitrate savings of 25% at 32kbits/s and of 33% at 24 kbits/s. The improvements are similar to those presented in [4], but their results concentrated only on the 24kbits/s case. Figure 2 illustrates the average grades for the considered methods in the listening tests. The conditions 'lp7000' and 'lp3500' are the hidden anchors of MUSHRA test, corresponding to low-pass versions of the original signal to 7kHz and 3.5kHz, respectively.

4. CONCLUSION

The present paper presents a new method for the quantization of the MDCT spectral coefficients of audio signals. It uses vector quantization in sub-bands and the parameterization of the quantization reso-

Method 1	Method 2	Gr.	Method 1	Method 2	Gr.
Cmpnd_128	AAC_128	0	Cmpnd_32	AAC_16	+1
Cmpnd_96	AAC_96	0	Cmpnd_24	AAC_32	0
Cmpnd_64	AAC_64	0	Cmpnd_24	AAC_24	+1
Cmpnd_64	AAC_64	0	Cmpnd_24	AAC_16	+1
Cmpnd_48	AAC_48	+1	Cmpnd_16	AAC_32	-1
Cmpnd_32	AAC_32	+1	Cmpnd_16	AAC_24	0
Cmpnd_32	AAC_24	+1	Cmpnd_16	AAC_16	+1

Table 4. Listening test results based on Student T-distribution with a confidence level of 95%. Grade values signify: '+1' first method is statistically better, '0' the two methods have statistically similar performance.

lution enables a successful use of a single method for bitrates ranging from 128kbits/s down to 16kbits/s. In addition to the improvements brought by the use of vector quantization, the success of the method is further guaranteed by a new optimization and encoding scheme of the spectral information. Furthermore, the use of several experimental observations corrects some deficiencies of the perceptual model within bit allocation algorithm. Listening tests on 44.1kHz mono files demonstrate similar performance to that obtained by the AAC for bitrates higher or equal to 64kbits/s and significant improvement for lower bitrates. The method can be further improved by considering the correlations between consecutive sub-bands, but there will be a pay-off in complexity.

5. REFERENCES

[1] ISO/IEC JTC1/SC29/WG11, "Information technology- Coding of audio-visual objects- Part 3: Audio," ISO/IEC 14496-3, 2001.

[2] A. Aggarwal, S. L. Regunathan, and K. Rose, "Asymptotically optimal scalable coding for minimum weighted mean square error," in *Proceedings of Data Compression Conference*, March, 27-29 2001, vol. 1, pp. 43–52.

[3] N. Iwakami, T. Moriya, and S. Miki, "High quality audio-coding at less than 64 kbit/s by using transform domain weighted interleave vector quantization (TWINVQ)," in *Proceedings of Intl. Conf. on Audio, Speech, and Signal Processing Conference*, May, 9-12 1995, vol. 5, pp. 3095–3098.

[4] N. Meine and B. Edler, "Improved quantization and lossless coding for subband audio coding," in *the 118th Convention of the Audio Engineering Society, Convention paper 6468*. Barcelone, Spain, May, 28-31 2005.

[5] R. Der, P. Kabal, and W.-Y. Chan, "Towards a new perceptual coding paradigm for audio signals," in *Proceedings of Intl. Conf. on Audio, Speech, and Signal Processing Conference*, 2003, vol. 5, pp. 457–460.

[6] F. Nordén and P. Hedelin, "Companded quantization of speech MDCT coefficients," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 163–173, March 2005.

[7] P. W. Moo and D. L. Neuhoff, "Optimal compressor functions for multidimensional companding," in *Proceedings of Intl. Symp. on Information Theory*, Ulm, Germany, June, 29 - July,4 1997, p. 515.

[8] ITU-R BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," 2003.

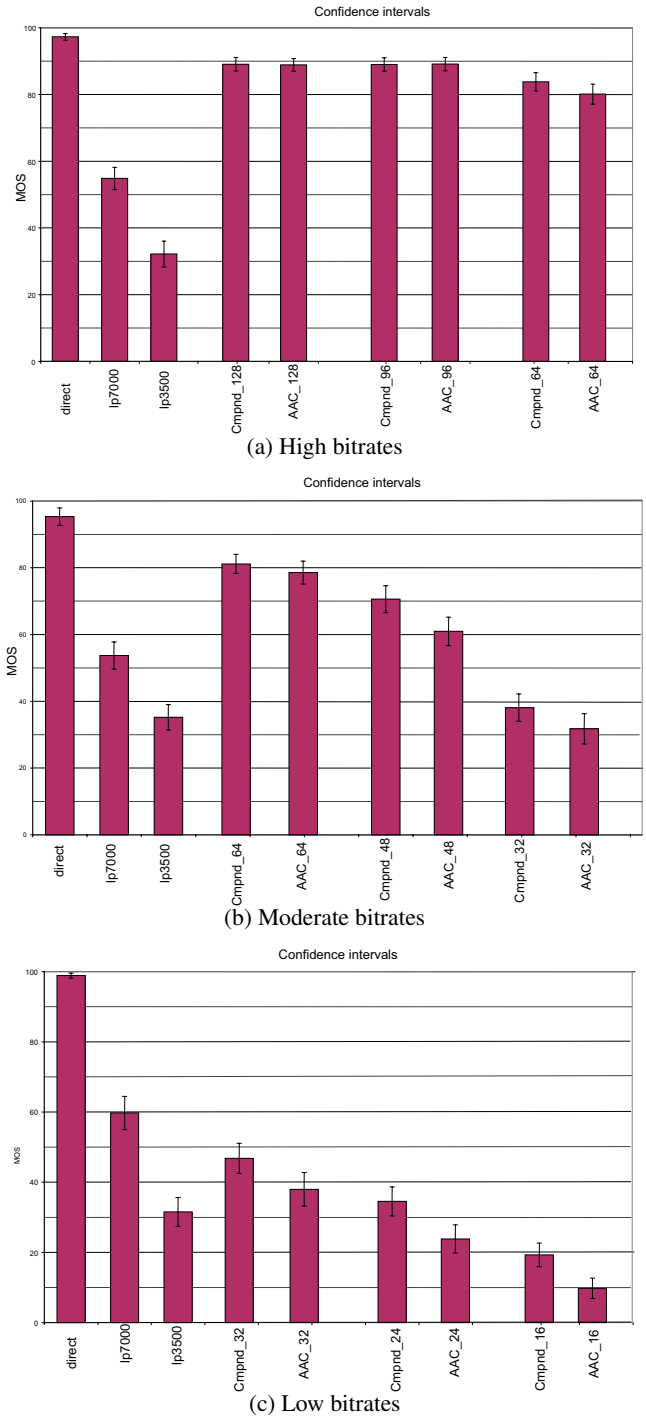

Fig. 2. Average of MOS values from listening tests.