

FINE-GRANULAR SCALABLE AND ERROR RESILIENT AUDIO CODING BY TREE-STRUCTURED QUANTIZATION

Stian Johansen Andrew Perkis[†] Tor A. Ramstad[†] Ajit S. Bopardikar

Quantifiable Quality of Service in Communications Systems
Norwegian University of Science and Technology
Trondheim, Norway
{stianjo, andrew, ajit}@q2s.ntnu.no

[†]Department of Telecommunications
Norwegian University of Science and Technology
Trondheim, Norway
tor@tele.ntnu.no

ABSTRACT

This paper¹² presents a scheme for scalable, error resilient coding of audio signals. The scheme is implemented here as part of a subband-based audio codec and involves the union of two interconnected ideas: Tree-structured vector quantization (TSVQ) and reallocation of bits (RoB). The former ensures that the encoding procedure results in fixed-length codes while the latter is a low-complexity scaling algorithm that computes optimal bitrates for each subband when the bitstream needs to be downsampled to a lower bitrate. The functionality enabled by these two systems assumes great importance when the encoded material is to be delivered over channels and networks that have a time-varying bitrate, such as IP-based wireless networks. Such channels are often error-prone and the potentially debilitating effects of introduced errors are mitigated by the inherent fixed-length property of all codewords. We present results that highlight the performance of this system under noisy conditions and show that the degradation in quality is graceful.

1. INTRODUCTION

Scalable source coding has been investigated by researchers for many years without being implemented in real-world systems. This fact is likely to change as broadcast and communicative services are being deployed onto a variety of terminals running over a variety of networks. Scalable coding can enable such a scenario by allowing for easy and cost-effective content adaptation and transcoding.

This paper presents a scheme for error resilient and fine-grained scalable audio coding. Unlike most standardized and proposed audio coding techniques, this scheme does not employ variable-length entropy coding. The use of such methods for data com-

paction inevitably leads to a varying bitrate (during transmission) which is susceptible to loss of synchronization in the case of channel errors. The presented scheme uses fixed-length coding and, furthermore, has a true fixed rate. That is, the instantaneous bitrate when transmitting will remain constant (which is not the case for variable length coding schemes). This is preferable in situations where the channel/network has an explicit upper limit for bandwidth usage.

A low-complexity algorithm for bitrate downscaling is presented, and it is shown that this has close to no loss of performance (in terms of SNR) when compared to direct encoding at the target rate. The properties of true fixed rate and error resilience do not change by employing such bitstream operations.

Our scheme consists of a subband coder using a 27-channel nonuniform tree-structured filterbank for signal decomposition. Subsequently, dynamic bit allocation based on rate-distortion measures is used along with tree-structured vector quantization. The scale factors are quantized using a fixed-length logarithmic quantizer.

This paper is organized as follows: Section 2 discusses the details of the developed audio codec. The algorithms and details of bit-rate scalability and error resilience are presented in sections 3 and 4, respectively. Results and concluding remarks are given in sections 5 and 6.

2. OVERVIEW OF THE CODEC

The basic structure of the encoder-decoder pair is shown in figure 1. The input PCM data is split into 27 subbands, and the needed segmentation and normalization is performed at the output of the filterbank. The normalization factors are quantized and used as input to the bit allocation algorithm. Based on these rate allocations, different tree-structured quantizers are used for the actual data compression. In the following, the elements of the developed scheme are treated in more detail.

2.1. Signal decomposition

A tree-structured nonuniform filterbank that tries to mimic the critical bands of the human ear [1] by 27 subbands is used. These bands represent a first order approximation of the ear's ability to separate sounds of different frequencies. The tree-structure is built using 2-channel perfect reconstruction filterbanks of the CQF (Conjugate Quadrature Filter) class, optimized with respect to coding gain. For this optimization, an AR(1)-process with an autocorrela-

¹This work was partly funded by "Centre for Quantifiable Quality of Service in Communication Systems, Centre of Excellence" appointed by The Research Council of Norway. <http://www.q2s.ntnu.no/>

²Copyright 2004 IEEE. Published in the 2004 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), scheduled for May 17-21, 2004 in Montreal, Quebec, Canada. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works, must be obtained from the IEEE. Contact: Manager, Copyrights and Permissions / IEEE Service Center / 445 Hoes Lane / P.O. Box 1331 / Piscataway, NJ 08855-1331, USA. Telephone: + Intl. 908-562-3966.

tion coefficient ρ of 0.95 is used for modelling the signal in question. This choice is motivated by the work presented in [2] and [3], where maximization of coding gain is shown to be of prime importance in filter design for coding of audio and images, respectively. The algorithm used for filter optimization is from [4]. The decision of what filter lengths to use at each split was made based on an exhaustive testing procedure (maximization of SNR). The filterbank structure is given in figure 3, where the filter lengths are given at the intersections.

2.2. Segmentation and normalization

The non-stationarity and dynamic range of audio signals is taken into account by segmentation and normalization after the signal has been decomposed into subbands. It is typically assumed that audio signals can be considered stationary over approximately 20ms [5]. At an input sampling rate of 44.1kHz, a segment length equivalent to 1024 input samples was chosen. This corresponds to a segment length of approximately 23ms. After taking the resulting downsampling factors of each band into account, the number of samples per segment per subband vary from 4 to 256 across subbands (see figure 3). Each of these segments is then normalized using the quantized sample standard deviation σ . The normalization with respect to these *scale factors* will make the *overall* subband probability density functions (pdfs) more Gaussian. During quantization, however, each segment is treated independently. This means that the pdf *shapes* input to the quantizer do not change; they are merely scaled.

2.3. Quantization of scale factors

The magnitude of the scale factors is varying with time and frequency, which makes design of appropriate quantizers difficult. The *logarithmic quantizer* approach outlined in [6] is used to overcome this. The representation points are given by

$$r_k = \beta \cdot \alpha^k, \quad (1)$$

where k is the quantization index and α and β are design parameters that must be chosen to match the quantizer to the dynamic range of the input signal. The normalization error is upper bounded by the following relation (second order approximation):

$$\text{Norm. error} \leq \frac{x \cdot \beta}{2\sigma^2} \cdot \left(\alpha \left\lceil \frac{\log \sigma}{\log \alpha} \right\rceil - \alpha \left\lfloor \frac{\log \sigma}{\log \alpha} \right\rfloor \right) \quad (2)$$

The implemented coder can use either 5 or 6 bits to represent k , while the parameter α varies across bands in order to account for different dynamic ranges. β is unity for all channels. This implies that the size of the side information (measured in bits/second) is *constant*, regardless of which rate the audio data is encoded at.

2.4. Bit allocation and quantization

The approach taken for bit allocation is the greedy algorithm based on optimal bit allocation (from [7]):

$$b_i = \frac{B}{M} + \frac{1}{2} \log_2 \left(\frac{\sigma_{y_i}^2}{\sigma_{y_g}^2} \right), \quad i = 0, 1, \dots, M-1 \quad (3)$$

where b_i is the optimal number of bits for component number i , M is the number of components (subbands), B is the total number of bits to use for quantizing and $\sigma_{y_g}^2$ is the geometric mean of

component variances. The greedy algorithm ensures that the bit allocation step results in rates that are positive integers (this is not always the case for equation (3)). It should also be noted that (3) is a high-rate approximation derived for simultaneous *scalar* quantization of subbands when the resulting downsampling factors are equal. This makes it somewhat inaccurate for the architecture presented here. Implementing psychoacoustic models along with the optimal bit allocation presented in [8] should give better results, but extensions to vector quantization are needed.

2.5. Quantization

The developed coder uses Tree-Structured (Vector) Quantization [9], a special case of vector quantizers that exists in different variants. We propose the use of fixed-length codes (balanced quantizer trees), since this allows for error resilience and a directly scalable bitstream. These tree-structured quantizers additionally have the desirable property of being computationally efficient. For a codebook size N , only $2 \cdot \log_2 N$ distance comparisons have to be computed when encoding, compared to N for general VQ. Based on the rates decided by the bit allocation algorithm, the quantizers to be used are chosen from the library of available quantizers. The quantizer with the highest dimensionality available for the desired rate is used. In the current implementation, quantizers of dimensions 8, 4, 2 and 1 are available. Designing one-dimensional (scalar) quantizers was needed because the computational complexity of designing high-rate multidimensional quantizers is very high. However, these scalar quantizers are still tree-structured, and thereby not violating the desired properties of scalability and error resilience.

3. BIT-RATE SCALABILITY

The use of tree-structured quantization and no variable-length entropy coding techniques allow for computationally efficient and accurate bit-rate scalability. Quantizer codewords can simply be truncated — this is equivalent to stopping decoding at a certain depth of the quantizer tree. The important question is: What would be the optimal depth to stop decoding, given a certain target bit rate? The algorithm presented in the following section describes a method for estimating this. It can be seen that this algorithm is optimal with respect to the bit allocation algorithm used.

3.1. ‘Reallocation of bits’(RoB)-algorithm

For close-to-optimal bitrate downscaling, the bit distribution has to be recalculated. Running the same bit allocation algorithm that was used by the encoder a second time with the target rate as input will yield the optimal use of bits for this new rate. The procedure is outlined in figure 2.

Note that the second pass must be restricted to using the same quantizers that were used during encoding. A comparison of the original and target rates for each subband/timeframe will decide what the resulting tree depth is. Since the tree depth will be reduced by the same fraction for the entire segment (this is a direct consequence of the bit allocation algorithm used), the resulting codewords will still inherit the desirable fixed-length property. Some results on the SNR loss incurred by using this algorithm are shown in Table 1. It is evident that the loss compared to direct encoding at the target rate is negligible (in terms of SNR). Since

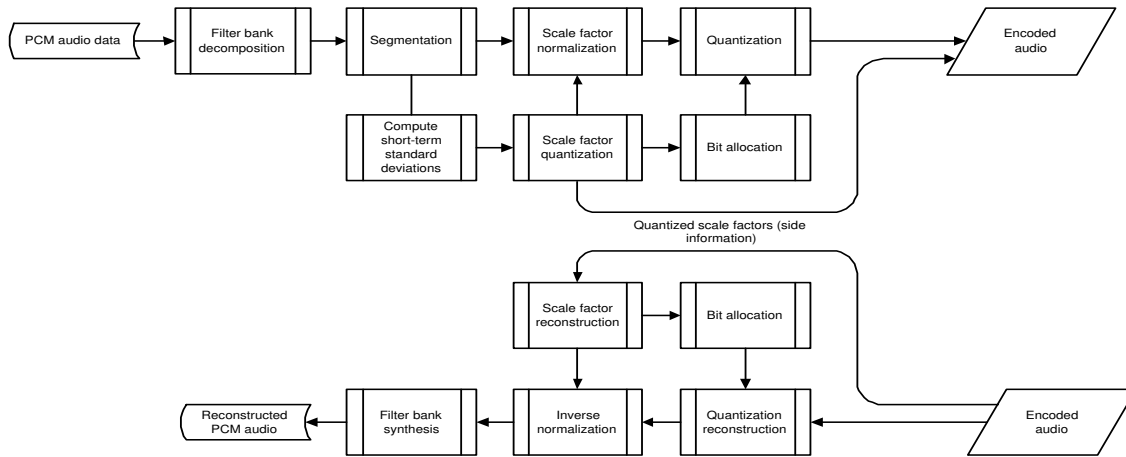


Fig. 1. Flowchart of the encoder/decoder

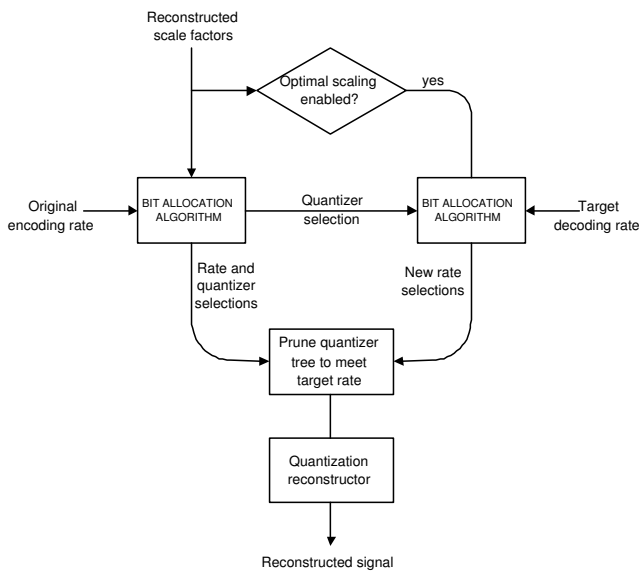


Fig. 2. Flowchart for the 'Reallocation of bits' (RoB) algorithm

| Target rate [kbps] | Direct SNR [dB] | Downscaled SNR [dB] | Δ [dB] |
|--------------------|-----------------|---------------------|---------------|
| 70 | 26.012 | 25.732 | -0.280 |
| 60 | 25.685 | 25.283 | -0.402 |
| 50 | 24.951 | 24.589 | -0.362 |
| 40 | 24.107 | 23.734 | -0.373 |

Table 1. SNR losses (Δ) when downscaling to the given rates compared to encoding directly at these rates. The downscaled data was encoded at 90kbps [mono].

the bit allocation algorithm is used when generating the downsampled bitstream, this scheme can meet *any* desired target rate. This separates our scheme from traditional differentially encoded layer-based coders. The recent MPEG-4 Audio standard implements Fine-Granular Scalability (FGS) [10] in steps of approxi-

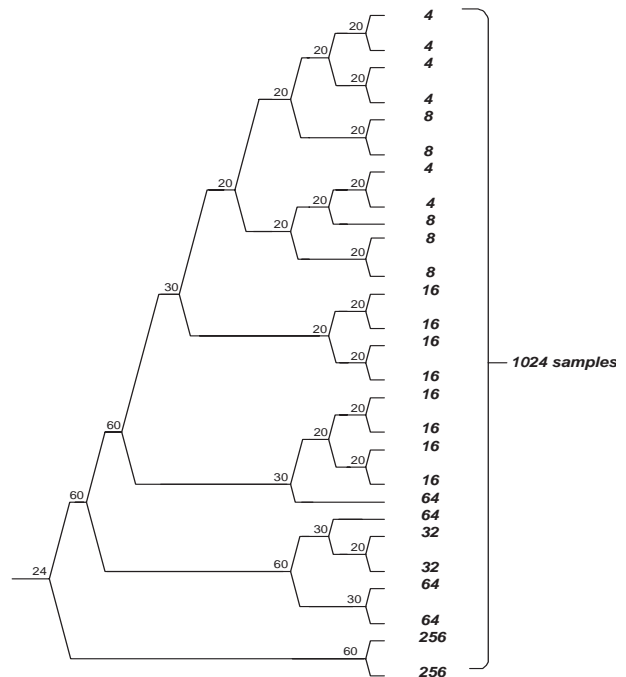


Fig. 3. Structure of the filterbank. The number of samples for each segment (and total) for the different channels is given on the right-hand side.

mately 1kbps/channel. However, in MPEG-4 this functionality is achieved through bit-plane coding and has an operating range of 16 to 64 kbps/channel. The proposed scheme has no such limitations.

4. ERROR RESILIENCE

Coding methods that employ variable-length entropy coding techniques are subject to loss of synchronization in the presence of residual channel errors. This is usually mitigated in one of two ways: Insertion of resynchronization markers or use of reversible variable length coding (RVLC). The latter was recently proposed

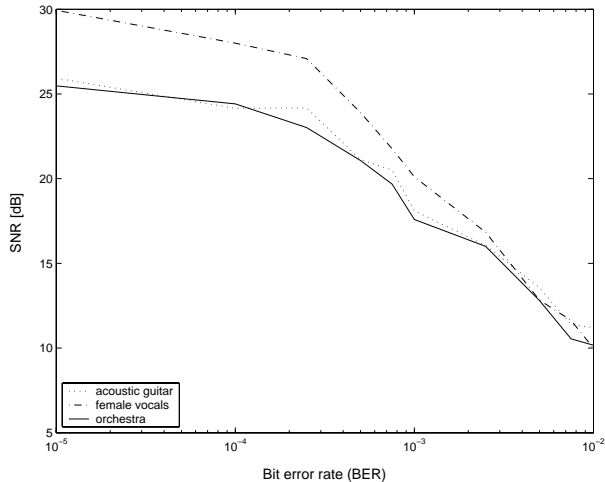


Fig. 4. SNR as a function of Bit Error Rate (BER) for three test sequences. The audio data was encoded at 80 kbps.

for audio coding in a paper by Zhou *et al.*[11]. However, the increased overhead can be considerable if error resilience is important (this is especially true in the case of resynchronization markers). The scheme presented here uses exclusively fixed-length codewords, both for the quantized data and the side information. This has the fortunate effect of not increasing overhead, while at the same time rendering loss of synchronization impossible. Figure 4 summarizes simulation results when bit errors are introduced in the encoded bitstream for three different audio clips. The results show that the incurred SNR loss is moderate at bit error rates around 10^{-4} ; this statement is also true for the perceived quality. At higher error rates the introduced distortion is indeed audible, which is to be expected. Most importantly, decoding is possible at *any* bit error rate. This obviously does not hold when errors are introduced in the header information. In this case, the decoder would fail because erroneous bit allocation tables would be produced.

5. RESULTS AND LISTENING TESTS

The main aim of this work has been to develop an audio coder that is scalable and error resilient. The implementation has been tested both for SNR and perceptual performance. Figure 5 shows the SNR performance of the developed coder as a function of bitrate (for monophonic audio). Informal listening tests have also been carried out; these indicate transparent performance at rates of 80kbps and above. These tests also show a certain 'crystallization' of the sound at low rates. This stems from the fact that no psychoacoustic models are currently implemented — bit allocation is done based on a rate-distortion measure only. This will favour the high-energy lower frequency bands at low rates while sacrificing clarity and crispness in the higher frequency bands.

6. CONCLUSIONS

The coding scheme presented herein has an inherent structure that allows for low-complexity bitrate scaling with a marginal loss compared to re-encoding at the target rate. The scaling algorithm that attains this performance is able to meet any desired rate. Inclu-

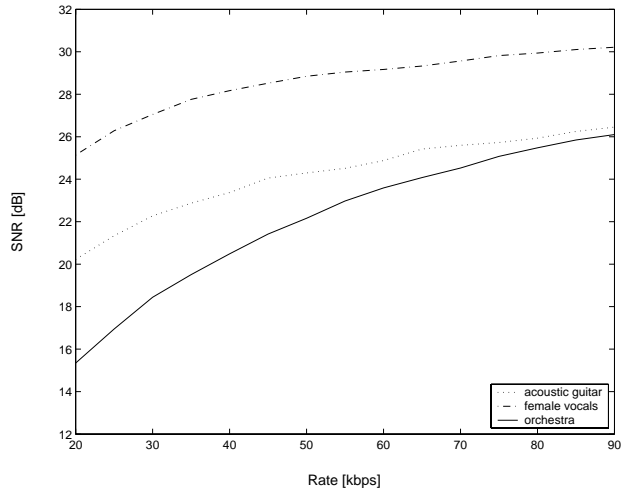


Fig. 5. SNR performance for the developed coder for the three test sequences in figure 4 (monophonic audio).

sion of psychoacoustic models, a more sophisticated bit allocation algorithm and a larger quantizer library in this scheme will considerably increase efficiency while still maintaining the useful properties discussed above.

7. REFERENCES

- [1] E. Zwicker, *Psychoakustik*, Springer-Verlag, Berlin, 1982.
- [2] P. Philippe, F. Moreau de Saint-Martin, and L. Mainard, "On the Choice of Wavelet Filters for Audio Compression," in *Proc. ICASSP'95*, Detroit, USA, 1995, vol. 2, p. 1045.
- [3] P. Onno and C. Guillemot, "Tradeoffs in the design of wavelet filters for image compression," in *Proceedings of SPIE; Visual Communications and Image Processing*, Boston, 1993, vol. 2094.
- [4] Y. Liu H. Caglar and A. N. Akansu, "Statistically optimized PR-QMF design," in *Proc. SPIE Visual Comms. and Image Processing*, 1991, vol. 1605.
- [5] T. Painter and A. Spanias, "Perceptual coding of digital audio," in *Proceedings of the IEEE*, 2000, vol. 88.
- [6] T. A. Ramstad, "Considerations on quantization and dynamic bit allocation in subband coders," in *Proc. ICASSP'86*, Tokyo, Japan, 1986.
- [7] T. A. Ramstad, S. O. Aase and J. H. Husøy, *Subband Compression of Images: Principles and Examples*, Elsevier Science B.V., Amsterdam, 1995.
- [8] X. Wei, M. J. Shaw and M. R. Varley, "Optimum bit allocation and decomposition for high quality audio coding," in *Proc. ICASSP'97*, 1997.
- [9] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1992.
- [10] S.-H. Park S.-W. Kim and Y.-B. Kim, "Fine grain scalability in MPEG-4 audio," in *Convention paper 5491, 111th AES convention*, New York, USA, 2001.
- [11] J. Zhou, Q. Zhang, Z. Xiong and W. Zhu, "Error resilient scalable audio coding (ERSAC) for mobile applications," in *Proc. IEEE Fourth Workshop on Multimedia Signal Processing*, 2001.