SCALABLE AUDIO CODER BASED ON QUANTIZER UNITS OF MDCT COEFFICIENTS

Akio Jin[1], Takehiro Moriya[1], Takeshi Norimatsu[2], Mineo Tsushima[2], Tomokazu Ishikawa[2]

[1]NTT Human Interface Laboratories,
3-9-11, Midori-cho, Musashino-shi, Tokyo 180-8585, Japan E-mail: jin@splab.hil.ntt.co.jp
[2]Matsushita Electric Industrial Co., Ltd.
1006, Kadoma, Kadoma-shi, Osaka 571-8501, Japan E-mail: norima2@arl.drl.mei.co.jp

ABSTRACT

A scalable codec has been constructed by using transform coding and the basic modules for scalable encoder and decoder. It allows users to choose a variety of scalable configrations in the frequency domain. The basic module is a quantizer that can quantize MDCT (Modified DCT)[1] coefficients transformed from a variety of frequency regions. This module mainly works at bitrates of more than 8 kbit/s. We can also change the target frequency regions of the basic module's input-output signals in each transform frame; i.e., we can change the scalable structure according to the nature of input signals. In the scalable codec described here, the input-output signals are monaural and the sampling frequency is 24 kHz. The total bit rate of this scalable codec is more than 8 kbit/s. Subjective quality evaluation tests, mainly for musical sound sources, showed that it's sound quality is better than that of an MPEG2-layer3 codec at 8, 16, and 24 kbit/s when our scalable codec is construced of 8-kbit/s basic modules.

In combination with AAC (Advanced Audio Coding)[2], our scalable codec will be chosen as an international standard in ISO/IEC-MPEG-4/Audio.

1. INTRODUCTION

There are demands for a codec, that can work at any desired transmission bit rate or decoded frequency-band or computational complexity. One way to meet these demands is to use a scalable codec. On the internet, a scalable codec can encode audio signals or translate a bit stream while controlling the transmission-bit rate or frequency band of decoded sounds when the transmission-bit rate is changing from momoent to moment. Besides taking a long time to decode of high-quality sound from all of the scalable coded bit stream, we can also quickly decode low-quality sound by using some part of the same coded bit stream.

In this paper, we describe a scalable codec based on hierarchical quantization, which is achieved using some basic modules for the encoder or decoder. The basic module is mainly constructed from a TwinVQ (Transform-domain Weighted Interleave Vector Quantization) codec[3][4][5] which is a type of transform coding. Transform coder is used in audio coding[6]. This basic module is a quantizer for the MDCT coefficients, and it is used to obtain a desired hierarchical-quantization-arrangement with a simple algorithm. For example, we can make 4-layer-scalable codec as shown in Figure 1.

Figure 1 shows our proposed method. It uses four basic modules with input sampling frequency of 24 kHz. This is a four-layered scalable codec, but we can make any number of layers using these basic modules. The basic module for the first layer (#1) has a fixed range of input or output frequency. But, other basic modules (#2, #3, #4) have a variable range of input or output frequency with each frequency band width fixed. The frequency point information of each basic module is added to the coded bitstream. In this figure, for example, the input 4-kHz signal is quantized in #1 module first, and then its quantization error is quantized in #2 and #3 modules. We can change the width (frequency band width), height (number of bits for each frequency), and position (target frequency) of the each module rectangle in this figure. Since the basic module is used as a common module, we can get such a flexibility and low-complexity. Each basic module is expressed as a rectangle for convenience, but in fact, the bit allocation in each basic module is defined by the perceptual weighting.

In this paper, we give an overview of the TwinVQ codec, and explain the structure of the basic module for encoder or decoder, the structures of the scalable encoder or decoder. Finally, we report the results of subjective listening tests.

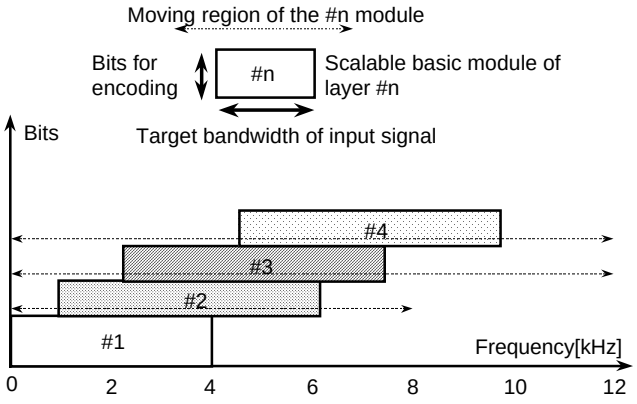

Figure 1. Hierarchical structure of scalable codec.

2. OVERVIEW OF TWINVQ CODEC

TwinVQ is an MDCT-based transform coder which has merits of high compression ratio and robustness against channel errors.[7] In the TwinVQ system (see Fig. 2), input signals are quantized by being transformed to MDCT coefficients. The length of the transform window is changed according to the nature of the input signals. In the encoder, the quantization of MDCT coefficients is done by LSP vector quantization, bark-scale-envelope vector quantization, gain quantization, and perceptual weighted interleave vector quantization of the prediction error. In the decoder, the inverse operations are carried out.

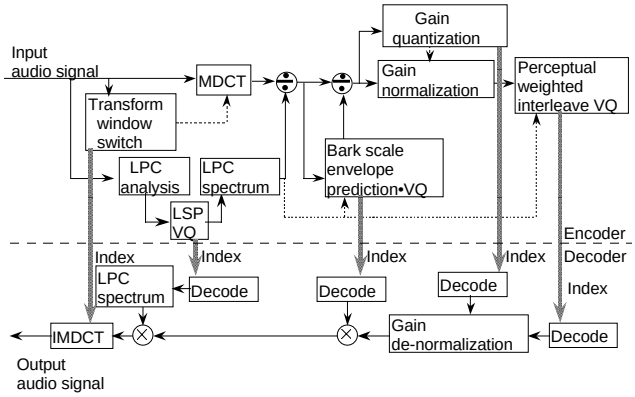

Figure 2. Block diagram of TwinVQ.

3. STRUCTURE OF BASIC MODULE FOR SCALABLE CODEC

In this scalable codec, we use the basic module for the scalable encoder or decoder as a common quantizer in each scalable layer. The aim is to carry out the hierarchical quantization of desired coefficients in the frequency region by a simple algorithm or simple layout. The figure 3 explains the structure of the basic module for the encoder. This structure looks like TwinVQ. The differences are discussed below.

- Input signals are MDCT coefficients, and not time-series signals.
- LPC is carried out not from time-series signals, but from MDCT coefficients. (Fig. 4) This is because we have contrived to aviod carring out MDCT transformation and inverse transformation repeatedly in order to simplify the scalable structure. The “previous method” in Figure 4 is the method used in TwinVQ.
- The basic module needs input information about the target frequency region. This information enables perceptual weighted vector quantization to be carried out in that target frequency.

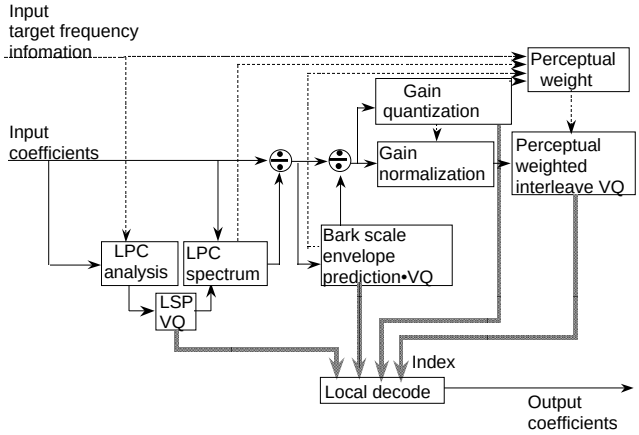

Figure 3. Block diagram of basic module for scalable encoder.

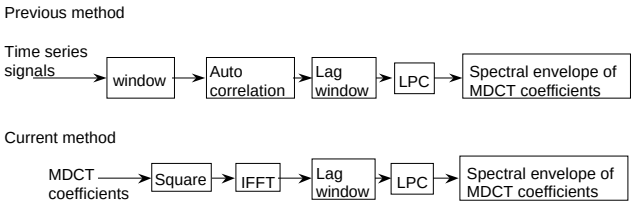

Figure 4. Block diagram of LPC from MDCT coefficients.

4. STRUCTURE OF SCALABLE ENCODER

4.1 ENCODER SYSTEM

The scalable encoder (Fig. 5) has parts for MDCT, hierarchical quantization, and bit stream generation. In the MDCT, the input time-series signals are transformed to MDCT coefficients according to its nature by particular transform points.

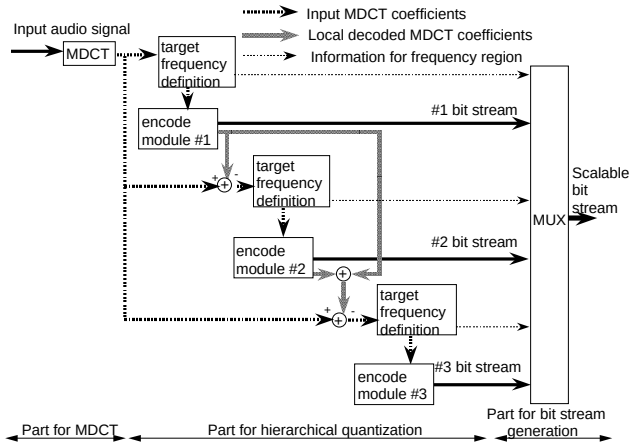

Figure 5. Structure of scalable encoder.

In the hierarchical quantization part, the input MDCT coefficients are quantized hierarchically by some basic modules for encoder. In this case, in the target frequency definition part, the target frequency region of each basic module is defined and coded as input signals of the basic module. In this paper, two bits are assigned for every frame and every layer to select the quantizing frequency position from 4 predetermined positions. In the bit stream generation part, the scalable bitstream is generated by lining up the codes from each basic modules.

4.2 TARGET FREQUENCY DEFINITION OF THE BASIC MODULE

In Figure 5, each basic module needs information about the frequency position. The position is defined by calculating the maximum position of the input coefficient power. In this paper, if that information is not given, the default position is used. The most typical positions are shown in the rectangles in Figure 1.

4.3 VQ CODEBOOKS

In the basic modules for encoder or decoder, we use some VQ codebooks. The types of codebooks are shown in Table 1. There are two type of codebook for each scalable layer, and four for each VQ. Amang these, there are two type of Shape VQ (Interleave VQ) codebook (for short frames or long frames). The codebooks for layer #2 and above are trained by signals in a variety of frequency positions and scalable layers. That is why, the codebooks for layer #2 and above are non-specialized ones.

Table 1. Codebook patterns for the basic modules.

Layer NO.	LSP VQ	Bark VQ	Shape VQ
layer #1	C_{lsp1}	C_{bark1}	$C_{shape\text{-}S1}$, $C_{shape\text{-}L1}$
layer #2 and above	C_{lsp2}	C_{bark2}	$C_{shape\text{-}S2}$, $C_{shape\text{-}L2}$

4.4 BIT ASSIGNMENT

Example of the bit assignments are listed in Table 2. In this table, the bit assignment of #1 module is sligitly diffrerent from the #2 module or above one. And some bits are reserved for the framework of the transform coding mode of MPEG-4/Audio.

5. STRUCTURE OF SCALABLE DECODER

The scalable decoder (Fig.6) has three parts for analizing the bit stream, hierarchical requantization, and IMDCT. The signal processing follows the inverse order to that of the encoder. You can select the quality of decoded sound by selecting the combination of the basic modules for the decoder, that is to say, by selecting the switch pattern in Figure 6.

Table 2. Example of bit assignment.

(L/Long frame, S/Short frame)

	#1 module	**#2 module and above**
Bit rate [kbit/s]	8	8
Sampling [kHz]	24	24
Frame size [point]	960	960
Short frame size [point]	120	120
Frame gain (L/S) [bit/fr]	9/41	8/40
LPC order	20	20
LSP quantization [bit/fr]	19	19
Prediction flag [bit/fr]	1	1
1st-stage VQ [bit/fr]	6	6
2nd-stage VQ [bit/fr]	4	4
Number of split at 2nd-stage	3	3
Bark-scale envelope (L/S) [bit/fr]	43/0	43/0
Band limitation switch [bit/fr]	1	0
Postfilter switch (L/S) [bit/fr]	1/0	0/0
Pitch prediction switch (L/S) [bit/fr]	1/0	0/0
Window switch [bit/fr]	2	0
Additional switch [bit/fr]	2	0
Basic Module position bits [bit/fr]	0	2
Shape interleave VQ (L/S) [bit/fr]	242/256	248/260

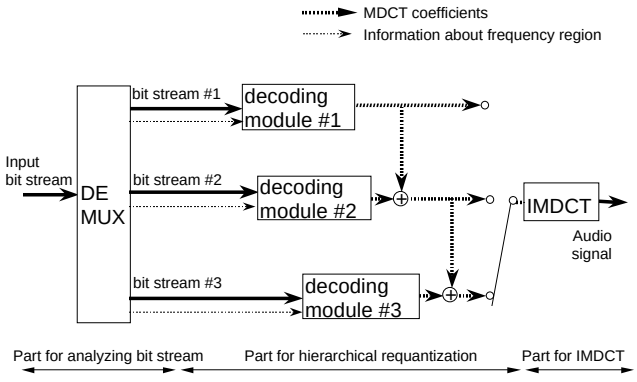

Figure 6. Structure of scalable decoder.

6. SUBJECTIVE LISTENING TESTS

Subjective listening tests were carried out in our laboratory to evaluate the quality of the scalable decoded sounds by Comparison Mean Opinion Score (CMOS) tests. For comparison, we tested an MPEG2-layer3 codec, which is not a scalable codec. There were 8 listeners, who were all involved in working with music. All the listeners were under 25. We used STAX-headphones, with 24-kHz sampling frequency of input-output signals, the signal type was monaural, and the length of MDCT transform window was 960 (Long) or 120 (Short) points. The number of sound sources was total 9 (1 speech, 6 instruments, 1 orchestra, and 1 big band). The sequence played to the listeners for each trial was Ref/A/B (Ref/B/A), where Ref was the original uncoded signal and A and B were both coded signals. The following 7-point grading scale was used by the subjects to compare the tested codecs. (Table 3.)

Table 3. 7-point grading scale for CMOS.

Score	Standard of estimation
+3	B is much better than A
+2	B is better than A
+1	B is slightly better than A
0	B is same as A
-1	B is slightly worse than A
-2	B is worse than A
-3	B is much worse than A

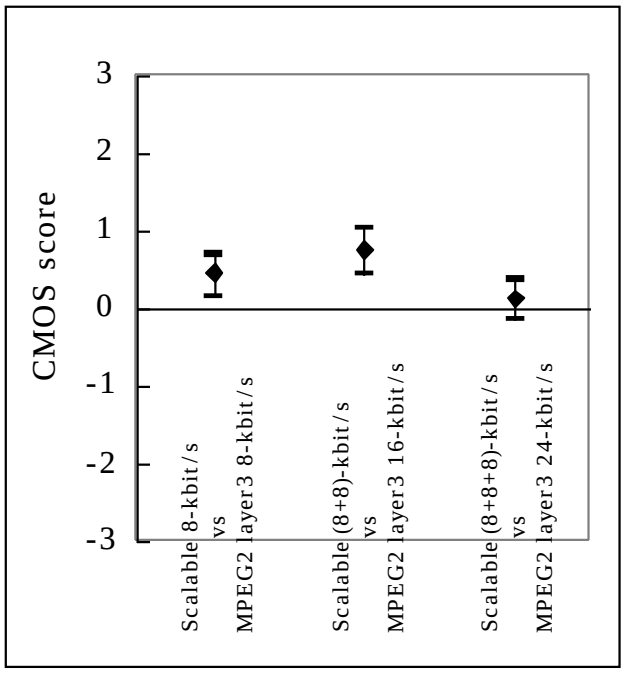

Figure 7. Result of OAB experiment.

The results CMOS are shown in Figure 7. Each value in Figure 7 is the mean score of 9 samples and 8 persons. And, the top and bottom values were upper and lower limits of the 95 percent confidence interval with respect to the sample mean. This figure shows that the 8-kbit/s scalable codec was significantly better than the 8-kbit/s (8-kHz sampling) MPEG2-layer3 codec, and that the 16-kbit/s (8 + 8 kbit/s) scalable codec was significantly better than the 16-kbit/s (16-kHz sampling) MPEG2-layer3 codec. The 24-kbit/s (8 + 8 + 8 kbit/s) scalable codec was slightly better than or equal to the 24-kbit/s (24-kHz sampling) MPEG2-layer3 codec.

7. SUMMARY

We made a scalable codec by which you can make your favorite hierarchical layout; i.e. that allows you to choose the quality level you desire. Its structure is simple as basic modules are used for encoder and decoder. Tests showed that the decoded sound quality of our scalable codec was better than or equal to that of the MPEG2-layer3 non-scalable codec.

8. REFERENCES

[1] J.P. Princen and A.B. Bradley, “Subband/transform coding using filter bank designs based on time domain aliasing cancellation”, Proc, IEEE ICASSP-87, pp.2161-2164, 1987.

[2] “ISO/IEC JTC1/SC29/WG11 13818-7 (MPEG-2 Advanced Audio Coding, AAC)”, 1997.

[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)”, Proc. IEEE ICASSP-95, pp.3095-3098, 1995.

[4] T. Moriya, N. Iwakami, K. Ikeda, and S. Miki: “Extension and Complexity Reduction of TwinVQ Audio Coder”, Proc. IEEE ICASSP-96, pp.1029-1032, 1996.

[5] T. Moriya, N. Iwakami, A. Jin, K. Ikeda, and S. Miki: “A Design of Transform Coder for Both Speech and Audio Signals at 1 bit/sample”, Proc. IEEE ICASSP-97, pp.1371-1374, 1997.

[6] K. Brandenburg and G. Stoll, “ISO-MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio”, J. Audio Eng. Soc., vol. 42, No. 10, pp. 780-792, 1994.

[7] K. Ikeda, T. Moriya, N. Iwakami, and S. Miki: “Error protected TwinVQ audio coding at less than 64 kbit/s/ch”, Proc. 1995 Speech Coding Workshop, pp. 33-34, Sept. 1995.