An 8-32 kbit/s Scalable Wideband Coder Extended with MDCT-based Bandwidth Extension on top of a 6.8 kbit/s Narrowband CELP Coder

Masahiro Oshikiri [1], *Hiroyuki Ehara* [1], *Toshiyuki Morii* [1], *Tomofumi Yamanashi* [1], *Kaoru Satoh* [1], and *Koji Yoshida* [1]

[1] Next-Generation Mobile Communications Development Center, Matsushita Electric (Panasonic), Japan

{oshikiri.masahiro,ehara.hiroyuki,morii.to,yamanashi.tomofumi,
satoh.kaoru,yoshida.koji}@jp.panasonic.com

10.21437/Interspeech.2007-478

Abstract

In this paper, we present a 6.8-32 kbit/s scalable speech and audio coder using a modified-discrete-cosine-transform (MDCT)-based bandwidth extension on top of a 6.8 kbit/s code-excited-linear-prediction (CELP) coder. The proposed coder comprises a 6.8 kbit/s narrowband CELP as its core-layer and eight enhancement layers with the bitrates of 0.8, 1.2, 3.2, or 4.0 kbit/s. After encoding of a narrowband signal by the core-layer, the first enhancement layer extends the bandwidth of a narrowband decoded signal, and the other enhancement layers increase the fidelity of an extended wideband signal or robustness against frame erasure conditions. Subjective evaluation test results demonstrate that the proposed coder outperforms G.729.1 for music signals at 16 and 24 kbit/s in particular with competitive or even better performance in other conditions like clean speech, background noise, and frame erasure.

Index Terms: speech coding, audio coding, scalable coding, bandwidth extension, MDCT, CELP, vector quantization

1. Introduction

Recently, several standardization activities related with scalable speech and audio coding have been launched in the International Telecommunication Union – Telecommunication Standardization Sector (ITU-T). This is because there are needs of scalable coders for extending existing coders and for offering attractive features for the future heterogeneous networks. Those features include bitrate scalability, bandwidth scalability (support of wideband (7 kHz bandwidth) or super-wideband (15 kHz bandwidth) signal), and robust frame erasure concealment. In 2006, G.729.1, one of such scalable coders, was standardized. Main features of G.729.1 [1] are; 1) a core layer bit-stream interoperable with G.729, 2) fine bitrate granularity with 12 layers, 3) support of wideband signals at the bitrate of 14 kbit/s or higher, 4) robustness against frame erasure by sending small redundant information.

On a parallel with the standardization of G.729.1, ITU-T Q.9/16 has continued to work on another scalable coder standard, tentatively named EV-VBR. While G.729.1 was standardized to satisfy an urgent need from certain market, EV-VBR is expected to be a totally new coder and to achieve better performance compared with G.729.1. Also, some new features will be provided, e.g. wideband capability is supported at the bitrate of 8 kbit/s. EV-VBR is scheduled to be approved in April 2008, and a baseline algorithm is selected in March 2007 by a competition. We participated in the competition and submitted a candidate algorithm in December 2006.

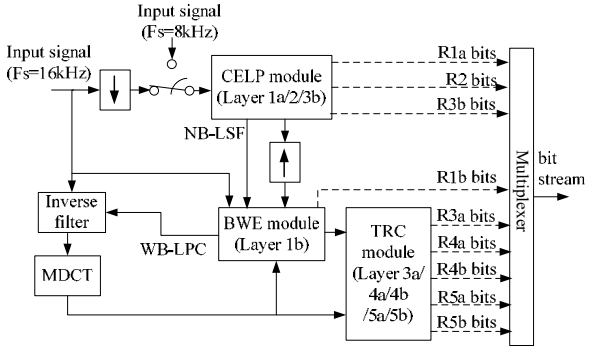

Figure 1: *Structure of the proposed encoder.*

This paper presents our proposed 6.8-32 kbit/s scalable coder. Rest of the paper organized as follows. The overview of the proposed coder is presented in Section 2. Descriptions of encoder and decoder are provided in Section 3 and 4 respectively. Results of our subjective evaluation test are discussed in Section 5. Finally, in Section 6, the paper concludes with remarks of the proposed coder.

2. Overview of the proposed coder

Our proposed coder operates on 20 ms frame and supports 8 and 16 kHz sampled inputs. The coder comprises a 6.8 kbit/s core layer and eight enhancement layers with the bitrate of 0.8, 1.2, 3.2 or 4 kbit/s, and therefore the bit stream can be truncated at the points of 6.8 kbit/s, 8 kbit/s, 12 kbit/s, 15.2 kbit/s, 16 kbit/s, 20 kbit/s, 24 kbit/s, and 28 kbit/s.

The coder has three main modules; a code-excited-linear-prediction (CELP) module, a bandwidth-extension (BWE) module and a transform-coding (TRC) module. The CELP module is designed to encode narrowband signal with high efficiency. Accompanied with the BWE module, the CELP module becomes capable to work as a wideband coder. The BWE module extends the bandwidth from narrowband to wideband in frequency domain using the modified-discrete-cosine-transform (MDCT). The TRC module encodes the fine spectral structure of wideband signal using an MDCT-based coder on top of the CELP and BWE modules. In addition to the three main modules, the coder includes a technique for encoding parameters used in frame erasure concealment (FEC) to enhance robustness against frame erasure conditions.

3. Encoder

A block diagram of the encoder is shown in Figure 1. A 16 kHz sampled wideband signal is inputted to a down-sampler and the down-sampler converts the input signal to an 8 kHz sampled narrowband signal. The narrowband signal is fed to a

CELP module that outputs R1a, R2 and R3b bits. The R1a and R2 bits are generated by a cascaded CELP encoder in the CELP module. The cascaded CELP has two layers, a core layer and an enhancement layer, and they generate the R1a bits and R2 bits respectively. The R3b bits are generated by an FEC encoder in the CELP module.

A local-decoded signal of the CELP module is upsampled to a 16 kHz sampled signal. The upsampled signal is provided to the BWE module. Line spectral frequencies (LSF) quantized by the CELP module and the 16 kHz sampled input signal is also inputted to the BWE module.

In the BWE module, a spectral envelope component and a spectrum fine structure component of the local-decoded signal are separately extended to wideband from narrowband. The spectral envelope component is represented with LSF, and the spectrum fine structure component is represented with the spectrum of a linear prediction (LP) residual signal. The extension of the LSF is realized by a wideband LSF (WB-LSF) quantization based on a predictive quantization using quantized narrowband LSF (NB-LSF) [2]. The extension of the LP residual spectrum is performed by a kind of pitch filtering in a frequency domain [3]. The BWE module produces R1b bits comprising extension information on both of the spectral envelope component and the spectrum fine structure component. The BWE module outputs quantized LP coefficients and an extended spectrum of the LP residual signal to an inverse filter and the TRC module respectively.

The inverse filter generates an LP residual signal of the 16 kHz sampled input signal and provides it to an MDCT section. The MDCT is performed every 40 ms windowed signal with 50 % overlap.

The TRC module encodes the errors between the MDCT coefficients obtained from the input signal and the signal decoded by the BWE module with 64 (R3a) or 80 (R4a, R4b, R5a and R5b) bits steps by using a gain-shape vector quantization.

Bit stream structure of the proposed coder is shown in Figure 2, and information on bit allocation is shown in Table 1 respectively.

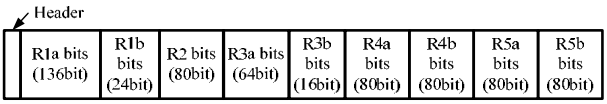

Figure 2: *Bit stream structure.*

Table 1. *Bit allocation of the proposed coder.*

Layer	Type	Parameter	Number of bits
1a	CELP	NB-LSF	21
		ACB	7.5+4+7.5+4
		FCB	16x4
		Gains	7x4
	Subtotal		136
1b	BWE	WB-LSF	11
		Mod. factor	4
		Pitch lag	4
		Gains	5
	Subtotal		24
2	CELP	FCB	14x4
		Gains	6x4
	Subtotal		80
3a	TRC	Region	3
		Shape	50
		Gains	10
		Reserved	1
	Subtotal		64
3b	FEC	Pulse pos.	7
		Signs	1
		Gains	8
	Subtotal		16
4a/ 4b/ 5a/ 5b	TRC	Region	3
		Shape	60
		Gains	16
		Reserved	1
	Subtotal		80
Total			640

3.1. CELP module

The CELP module processes an 8 kHz sampled narrowband signal. This module includes a cascaded CELP coder having two layers, Layer-1a and Layer-2. An FEC encoder is also included in this module for encoding the information served for frame erasure concealment.

Layer-1a encodes the input signal with the bitrate of 6.8 kbit/s. An error signal between the input signal and a signal decoded locally by Layer-1a becomes a target for Layer-2 encoding with the bitrate of 4 kbit/s.

In Layer-1a, a 12-order LP analysis is performed on every 20 ms frame. After converting to LSF, 12 LP coefficients are quantized in the LSF domain. The NB-LSF is quantized by an MA-predictive two-stage vector quantizer. Excitation parameters, adaptive codebook (ACB), fixed codebook (FCB) and gain parameters, are encoded on every 5 ms subframe so that perceptually weighted encoding distortion becomes minimum. The perceptual weighting is performed by a typical pole-zero filter but having an adaptive tilt control function. Differential quantization is exploited to quantize the ACB lags for even subframes with using 4 bits. The lags for odd subframes are combined and quantized with using 15 bits. This results in the 7.5 bits assignment for the first and third subframes. The FCB is a multi-dispersed-pulse based codebook, which generates an FCB vector composed of one or more dispersed pulses. 16 bits are used for encoding combination of pulses and their positions. A pair of ACB and FCB gains is quantized by a 7-bit MA-predictive vector quantizer. The FCB vector and the gain pair are quasi-simultaneously optimized using a joint codebook search approach.

The Layer-2 has a similar structure of the Layer-1a. That is, the encoding error of Layer-1a is further quantized by another CELP layer. However, the Layer-2 utilizes the LSF and ACB lag quantized by the Layer-1a as Layer-2 parameters, and therefore an FCB vector and a gain pair are the parameters encoded in the Layer-2. The FCB vector is encoded with 14 bits, and the gain pair is encoded with 6 bits.

The FEC encoder quantizes the following parameters obtained in the previous frame; the position, sign and amplitude of the last pitch pulse, the energy of the excitation signal, and the prediction gain of the LP filter. The pitch pulse amplitude, excitation energy, and the filter prediction gain are jointly vector-quantized in a logarithmic scale with using 8 bits. 7 bits and 1 bit are assigned respectively to encode the position and sign of the pitch pulse.

3.2. BWE module

A block diagram of the BWE module is shown in Figure 3. The BWE module extends the bandwidth of the local-decoded narrowband signal from the CELP module to wideband. In BWE module, extension processing is performed separately on a spectral envelope component and a spectrum fine structure component. The spectral envelope component is quantized as LSF, and the spectrum fine structure component is encoded as an LP residual spectrum.

A WB-LSF encoder is used to extend the NB-LSF quantized by the CELP module into WB-LSF. 11 bits are used by the WB-LSF encoder for quantizing the inputted WB-LSF with predictive vector quantization in which the quantized NB-LSF is exploited efficiently for predicting the WB-LSF. A quantized WB-LSF is converted to LP coefficients (WB-LPC), which are given to two inverse filters.

The inverse filter in the BWE module performs inverse filtering on the upsampled signal decoded by the Layer-2 to flatten its spectrum. The resultant signal is transformed into MDCT coefficients. At this stage, the valid bandwidth of the MDCT coefficients is still narrowband.

A WB-EXC encoder performs to extend the narrowband MDCT coefficients into wideband MDCT coefficients. A high frequency band of the original MDCT coefficients, which are obtained by applying the inverse filter to the 16 kHz sampled input signal, are approximated using the narrowband MDCT coefficients. The approximation algorithm has two main steps; a step of modification to the narrowband MDCT coefficients and a step of pitch filtering on the narrowband MDCT coefficients. The modification step is aimed at fitting the dynamic range of the narrowband MDCT coefficients to that of the high frequency band coefficients of the original. 4 bits are assigned to encode the information for this step. In the next step, single- or multi-tap pitch filtering is performed on the modified narrowband MDCT coefficients to generate the high frequency band MDCT coefficients. The basic concept of this pitch filtering is to preserve a harmonics structure of the high frequency band spectrum by utilizing the narrowband spectrum. The parameters, a pitch lag and gains for the approximated MDCT coefficients, are sequentially encoded so that an encoding error is minimized. 4 and 5 bits are assigned to encode the lag and gains, respectively.

3.3. TRC module

The TRC module is a cascaded MDCT-based coder that has five layers; Layer-3a, Layer-4a, Layer-4b, Layer-5a and Layer-5b. Each layer has same structure as shown in Figure 4.

The target vector for each layer encoding is residual error between the original MDCT coefficients and the MDCT coefficients decoded by the layer prior to each layer. The target vector is divided into 17 uniform sub-bands. We defined eight regions which consist of five consecutive sub-bands. The eight regions are arranged to be overlapped with each neighbouring regions and to cover full-band of 7 kHz. Each layer of the TRC module encodes one of the eight regions. Using 3 bits, the region selection is done by referring the energy of the target vector in the region. Once the region is selected, the target vector in the region is quantized using a shape-gain quantization in which a shape and a gain of the target vector are sequentially quantized. Shape-vectors are formed from 5 main and 4 sub pulses or 5 main and 6 sub pulses, and the combination of those pulse positions and signs are encoded with 50 or 60 bits. Following to the shape

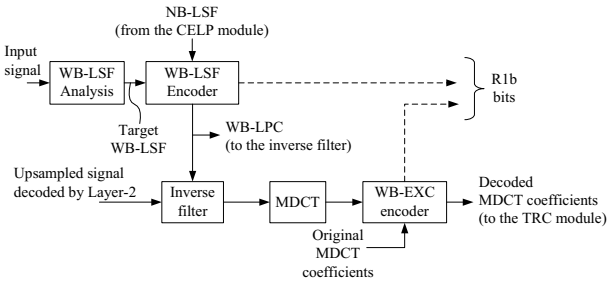

Figure 3: *Structure of the BWE module.*

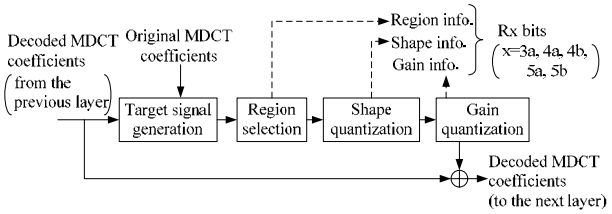

Figure 4: *Structure of the MDCT-based coder.*

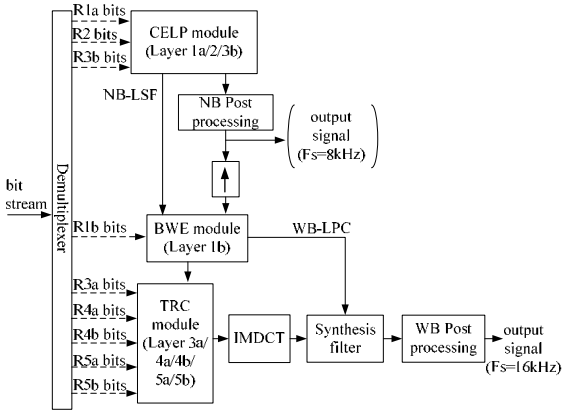

Figure 5: Structure of the proposed decoder.

quantization, a gain quantization is performed on calculated sub-band gains. Since the region contains 5 sub-band, 5 gains are obtained for the region and vector quantized using 10 or 16 bits. The vector quantization is exploits a switched prediction scheme.

4. Decoder

A block diagram of the decoder is shown in Figure 5. The input bit stream is de-multiplexed and the resultant bits are provided to corresponding modules i.e. CELP, BWE and TRC modules. There are two kinds of post-processing units in the decoder. One is an NB post-processing unit operated on an 8 kHz sampled signal, and the other is a WB post-processing unit operated on a 16 kHz sampled signal. Both of them have similar structure and comprise a pitch postfilter, a formant postfilter, a spectral tilt controller and a noise post-processor. These two post-processors do not activate simultaneously, but one of two is selectively enabled.

Since the input bit stream may be truncated, there are five possible configurations of decoding process depending on how many bits are received by the decoder. Each configuration is described below.

- **The case of receiving no bits**

In this case, a frame erasure concealment process is performed using the R3b bits received in the next frame or using the parameters decoded in the past. Since the next frame is supposed to be received and decoded for overlapping in the inverse MDCT operation, this procedure does not introduce any additional algorithmic delay.

- **The case of receiving R1a bits**

In this case, the CELP module and NB post-processing unit are activated. The CELP module generates a signal decoded by the Layer-1a using R1a bits. After passing through the NB post-processing unit and an up-sampler, the BWE module inserts zero values to high frequency band as decoded MDCT coefficients.

- **The case of receiving R1a to R1b bits**

In this case, the CELP and BWE modules and NB post-processing unit are activated. The BWE module obtains the upsampled signal decoded by the Layer-1a as described in the above case and extends the bandwidth of the signal to wideband by using parameters decoded from R1b bits.

- **The case of receiving R1a to R2 bits**

In this case, the CELP and BWE modules and NB post-processing unit are activated, too. But the CELP module generates a signal decoded by the Layer-2 using R1a bits and R2 bits. Therefore, the signal to be extended by the BWE module is the signal decoded by the Layer-2 after passing through the NB post-processing unit and the up-sampler.

- **The case of receiving R1a to R3a or higher layer bits**

In this case, all the modules except the NB post-processing unit are enabled. The CELP module generates the signal decoded by the Layer-2. After passing through the up-sampler, the decoded signal is extended to wideband through the BWE module. Then, the TRC module piles decoded residual MDCT coefficients on top of the extended Layer-2 MDCT coefficients.

5. Subjective evaluation tests

To assess the performance of the proposed coder, we conducted subjective listening tests using 16 non-expert listeners with monaural headphones. Absolute Category Rating (ACR) tests [4] were used for clean speech, frame erasure and music conditions, and Degradation Category Rating (DCR) tests [4] were used for background noise conditions. 8 Japanese sentence-pairs were used for clean speech and frame erasure conditions. For music conditions, 8 different audio sources, i.e. orchestra, opera, popular and jazz music, were used. For background noise conditions, we used 4 Japanese sentence-pairs corrupted with one of environmental noise sources, i.e. “Car noise” at 15 dB, “Street noise” at 20 dB, “Interference talker” at 15 dB or “Office noise” at 20 dB. We used G.729.1 as a reference coder. Test results are shown in Table 2. In the table, 95 % confidence intervals are given in parentheses, and the scores are written in boldface when the score of one coder indicates “statistically better than” that of the other coder.

From the test results, significant differences are found in 8 conditions. In 7 conditions out of the 8, the proposed coder outperforms G.729.1. Especially, the differences in the music condition at 16 and 24 kbit/s are distinctive. Although G.729.1 obtains better score in one condition, the bitrate of G.729.1 in the condition is higher than that of the proposed coder by 6 kbit/s. This is because G.729.1 does not support wideband inputs at the bitrate of 8 kbit/s.

Table 2. *Subjective evaluation test results.*

Bitrate and bandwidth	Condition	Proposed coder	G.729.1
8 kbit/s Narrowband	Clean speech	3.56 (0.15)	3.56 (0.13)
	3 % FER	3.55 (0.16)	3.44 (0.14)
	Car noise	**4.27** (0.19)	3.73 (0.23)
	Street noise	3.94 (0.23)	3.77 (0.25)
12 kbit/s Narrowband	Clean speech	3.72 (0.15)	3.58 (0.15)
	Car noise	4.58 (0.17)	4.45 (0.19)
	Street noise	4.16 (0.21)	4.27 (0.20)
8 kbit/s Wideband (G.729.1: 14 kbit/s)	Clean speech	3.48 (0.14)	3.37 (0.13)
	3 % FER	3.24 (0.14)	3.28 (0.11)
	Inter-talker	3.56 (0.21)	**3.89** (0.20)
	Office noise	4.03 (0.22)	4.25 (0.20)
16 kbit/s Wideband	Clean speech	3.44 (0.14)	3.37 (0.13)
	6 % FER	2.73 (0.16)	2.82 (0.14)
	Music	**2.91** (0.15)	2.22 (0.16)
	Inter-talker	4.09 (0.20)	3.97 (0.20)
	Office noise	4.39 (0.16)	4.25 (0.19)
24 kbit/s Wideband	Clean speech	3.53 (0.16)	3.43 (0.13)
	Music	**3.45** (0.15)	2.62 (0.17)
	Inter-talker	**4.52** (0.15)	4.30 (0.19)
	Office noise	4.61 (0.14)	4.53 (0.16)
32 kbit/s Wideband	Clean speech	**3.84** (0.14)	3.57 (0.14)
	3 % FER	3.40 (0.14)	3.35 (0.16)
	0-10 % FER	**3.82** (0.13)	3.53 (0.13)
	Music	**3.75** (0.14)	3.46 (0.15)
	Inter-talker	4.58 (0.16)	4.52 (0.18)
	Office noise	4.67 (0.13)	4.70 (0.14)

6. Conclusions

This paper presents an 8-32 kbit/s scalable wideband coder extended with MDCT-based bandwidth extension on top of a 6.8 kbit/s narrow band CELP coder. The proposed coder comprises a 6.8 kbit/s narrowband CELP as its core-layer and eight enhancement layers. A decoded narrowband signal is extended to a wideband signal using an MDCT-based bandwidth extension algorithm, and further enhancement is brought by enhancement layers of MDCT coders. Subjective test results demonstrated that the proposed coder outperforms G.729.1 for music signals at 16 and 24 kbit/s in particular with competitive or even better performance in other conditions, i.e. clean speech, background noise, and frame erasure conditions. The coder presented in this paper was submitted to the competition for ITU-T EV-VBR baseline algorithm selection.

7. References

[1] “G.729 based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream inter-operable with G.729,” ITU-T Recommendation G.729.1.

[2] H. Ehara, T. Morii, M. Oshikiri, K. Yoshida, and K. Honma, “Design of bandwidth scalable LSF quantization using interframe and intraframe prediction,” in proc. INTERSPEECH 2005, pp.1493-1496, 2005.

[3] M. Oshikiri, H. Ehara and K. Yoshida, “Efficient spectrum coding for super-wideband speech and its application to 7/10/15 kHz bandwidth scalable coders,” in proc. ICASSP 2004, pp.I-481-I-484, 2004.

[4] “Methods for subjective determination of transmission quality,” ITU-T Recommendation P.800.