IEICE TRANS. INF. & SYST., VOL.E93–D, NO.2 FEBRUARY 2010

LETTER

Delay-Reduced MDCT for Scalable Speech Codec with Cascaded Transforms

Hochong PARK[†a)], ***Member and*** **Ho-Sang SUNG**[††], ***Nonmember***

SUMMARY A scalable speech codec consisting of a harmonic codec as the core layer and MDCT-based transform codec as the enhancement layer is often required to provide both very low-rate core communication and fine granular scalability. This structure, however, has a serious drawback for practical use because a time delay caused by transform in each layer is accumulated, resulting in a long overall codec delay. In this letter, a new MDCT structure is proposed to reduce the overall codec delay by eliminating the accumulation of time delay by each transform. In the proposed structure, the time delay is first reduced by forcing two transforms to share a common look-ahead. The error components of MDCT caused by the look-ahead sharing are then analyzed and compensated in the decoder, resulting in perfect reconstruction. The proposed structure reduces the codec delay by the frame size, with an equivalent coding efficiency.

key words: *scalable speech codec, MDCT, transform codec, harmonic codec, time delay*

1. Introduction

Digital speech communications over packet-switched networks require a scalable speech codec for the fine adaptation of packet transmission to varying network capacities [1]–[4]. To provide a very-low-rate core layer and fine granularity in the enhancement layers, a scalable codec consisting of harmonic codec and MDCT (modified discrete cosine transform)-based transform codec for each layer is often desired.

A basic encoder structure of such a codec is depicted in Fig. 1. Each encoder block has a time-to-frequency transform module, and an overlap window with look-ahead is used in each transform to reduce the boundary effect between the adjacent frames. Consequently, each codec introduces a time delay equal to the amount of the current and the look-ahead regions, and all of the time delays are accumulated because of the cascaded connection of the multiple transforms, resulting in a long codec delay.

In this letter, a new MDCT structure in a transform codec is developed to reduce the overall codec delay by eliminating the accumulation of time delay by each transform. To eliminate delay accumulation, two transforms must share a common look-ahead region, but this condition in turn causes incorrect MDCT operation. To achieve both delay reduction and correct MDCT, a new idea is proposed in this letter, where the error components of MDCT due to a common look-ahead region are analyzed and compensated in the decoder, resulting in perfect reconstruction. Therefore, the proposed structure allows the codec shown in Fig. 1 to provide a very-low-rate core layer without increasing the delay, which is its main advantage over the CELP (code-excited linear prediction)-based scalable codec such as G.729.1 [4].

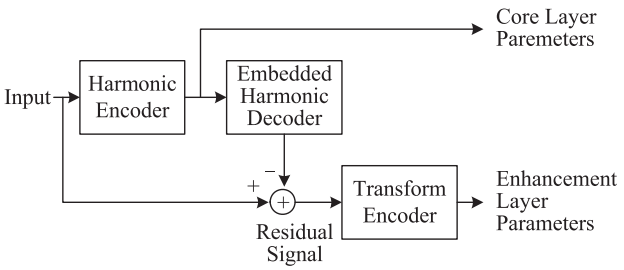

Fig. 1 Scalable codec with a serial connection of harmonic codec and transform codec.

The scope of this letter is limited to the development of a new MDCT structure with reduced time delay. Furthermore, this study does not include the operations of a harmonic codec and a transform codec including quantization and layer allocation. Various state-of-the-art techniques such as those in HVXC (harmonic vector excitation coding), AAC (advanced audio coding), BSAC (bit-sliced arithmetic coding), and G.729.1 can be used for this purpose [4]–[7].

2. Proposed Delay-Reduced MDCT Structure

2.1 Operation with Reduced Delay

It is assumed that the frame size is N and that both DFT (discrete Fourier transform) in harmonic codec and MDCT in transform codec use a 50% overlap window $w[n]$ of length $2N$ for maximum smoothing between the adjacent frames. It is also assumed that the window is symmetric, constant over the frames, and satisfies $w^2[n] + w^2[N-1-n] = 1$, $0 \le n < N$. These window conditions are necessary for the normal operation of MDCT [8].

Figure 2 shows the time diagram of codec operation without delay accumulation, where the current frame corresponds to the time index $n = 0, 1, \ldots, N-1$. Figure 2 (a) shows the original input signal and Fig. 2 (b) shows the IDFT (inverse DFT) output of core layer over two frames. In the normal scalable codec operation with delay accumulation, an overlapped-and-added output of the current frame

Manuscript received September 9, 2009.
Manuscript revised October 20, 2009.
[†]The author is with Kwangwoon University, Korea.
[††]The author is with Samsung Electronics, Korea.
a) E-mail: hcpark@kw.ac.kr
DOI: 10.1587/transinf.E93.D.388

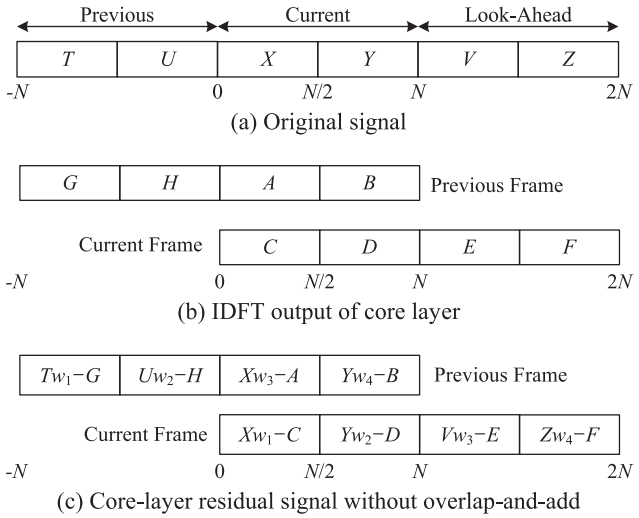

Fig. 2 Time diagram of the codec operation with the reduced time delay by sharing a common look-ahead region.

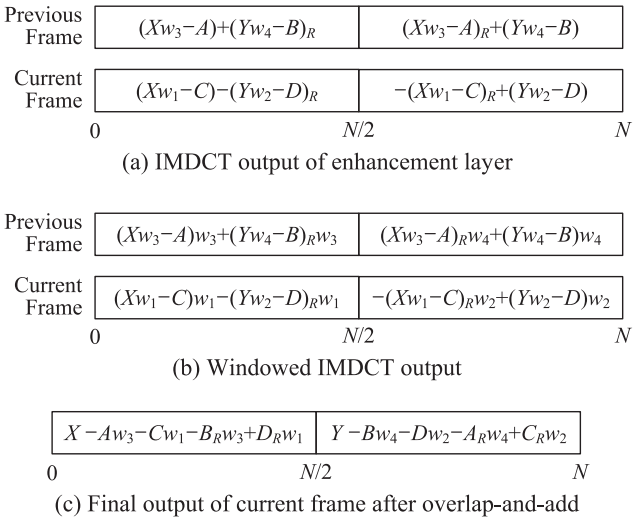

Fig. 3 Operations of MDCT and IMDCT with a common look-ahead region.

in the core layer is first computed using A, B, C and D. Then, a residual signal of length $2N$, $n = -N, -N+1, \ldots, N-1$, over previous and current frames is constructed and inputted into the current MDCT in the enhancement layer. In this case, an additional time delay N is added after IMDCT (inverse MDCT) and overlap-and-add operation, resulting in $3N$ overall codec delay.

In the proposed scalable codec structure, the delay accumulation can be eliminated by using the common look-ahead region for both layers, which can be realized by computing the residual signal without overlap-and-add operation after IDFT as in Fig. 2 (c). Since the window was applied before DFT in the core layer, the residual signal is expressed in the form of $[Xw_1 - C]$, where w_i is a piece of the window function of length $N/2$.

This residual signal of length $2N$ is then inputted into MDCT in the enhancement layer, which makes MDCT share the same look-ahead region with DFT in the core layer, resulting in an overall delay of only $2N$. Clearly, however, this structure causes incorrect MDCT operation because TDAC (time-domain aliasing cancellation) is impossible, and, as such, cannot be used in practice.

2.2 Analysis of Error Components

The codec operations shown in Fig. 2 including inherently incorrect MDCT operations are required for the purpose of delay reduction. Subsequently, the exact error components due to this unusual structure needs to be identified analytically and removed in the decoder, thereby eliminating all the incorrect operations in MDCT caused by a common look-ahead region.

When the residual signals in Fig. 2 (c) are inputted into MDCT, the IMDCT outputs in the overlap region of the previous and current frames become those in Fig. 3 (a) [8]. The subscript R indicates the time reversal within a given time slot of duration $N/2$. Since the residual signal was obtained without overlap-and-add operation, no window is applied to the residual signal before inputting it to MDCT. A window is applied only to the IMDCT output before overlap-and-add operation, resulting in the windowed IMDCT output of each frame as shown in Fig. 3 (b). The final output is obtained by overlap-and-add operation in the overlap region as shown in Fig. 3 (c) using the given window conditions. This output, however, is not correct due to look-ahead sharing when computing the residual signal.

To identify the error components, the target output of the enhancement layer is first determined. The output signal of the core layer in the decoder is an overlapped-and-added signal between the previous and current frames after windowing, which is equal to $[Aw_3 + Cw_1][Bw_4 + Dw_2]$, where $[P][Q]$ means a concatenation of two signals P and Q which have the same length of $N/2$. The final output of the decoder is the sum of the core- and enhancement-layer outputs, and the final output should be equal to the original signal $[X][Y]$; hence, the target output of the enhancement layer is $[X - (Aw_3 + Cw_1)][Y - (Bw_4 + Dw_2)]$. Comparing this with Fig. 3 (c), therefore, the error components in the given codec operations can be written in a compact closed form as $[-B_Rw_3 + D_Rw_1][-A_Rw_4 + C_Rw_2]$. If these components are removed from the overlapped-and-added signal after IMDCT in the decoder, the correct codec output can be obtained.

There is no need to transmit the error components from the encoder to the decoder because all the necessary values A_R, B_R, C_R, and D_R, are available in the decoder after the core layer. Hence, the proposed structure requires no additional bits and does not increase the bit rate at the expense of reduced time delay.

2.3 Performance Analysis

In the proposed structure, the residual signal and the resulting MDCT coefficients are modified due to a look-ahead

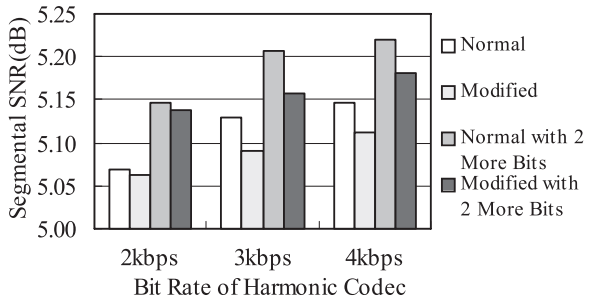

Fig. 4 Segmental SNR of 16 kbps quantization of MDCT coefficients.

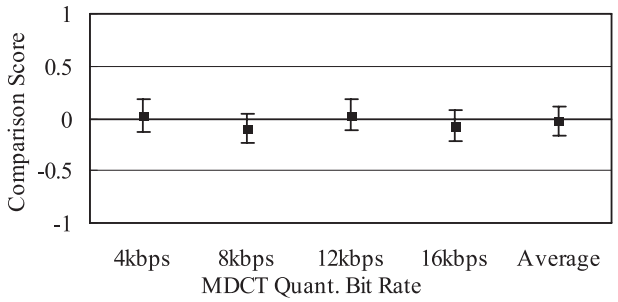

Fig. 5 Results of quality comparison between the normal G.729.1 and the modified G.729.1 with two more bits for MDCT quantization.

Table 1 The minimal number of additional bits per 20 msec-frame in order to make the modified coefficients secure the equivalent segmental SNR to the normal coefficients.

Bit Rate of Harmonic Codec(kbps)	2				3				4			
Bit Rate of MDCT Quant. (kbps)	4	8	12	16	4	8	12	16	4	8	12	16
Number of Additional Bits	0	0	0	1	0	1	2	2	0	0	1	1

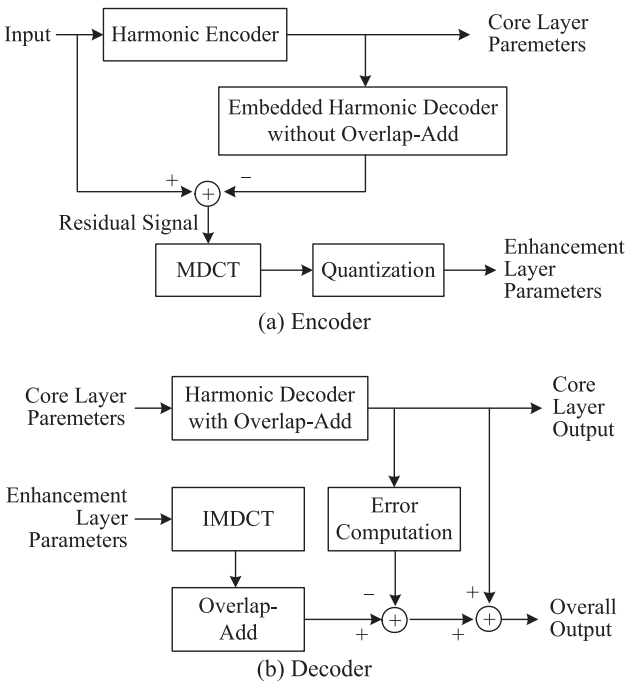

Fig. 6 Block diagrams of encoder and decoder equipped with the proposed MDCT structure.

sharing. Hence, such modification to the MDCT coefficients must be verified to not degrade the quantization performance, compared to that of the normal MDCT coefficients without a look-ahead sharing. For this purpose, using a prototype harmonic codec, the normal and the modified residual signals are computed, their respective MDCT coefficients are quantized using the quantizer of G.729.1, and their segmental SNR's (signal-to-noise ratios) are analyzed.

Figure 4 shows the segmental SNR for the 16 kbps quantization of MDCT coefficients. The modified coefficients have lower performance than the normal coefficients; however, when two more bits per 20 msec-frame are added to the quantization of modified coefficients, the segmental SNR becomes larger than that of normal coefficients as shown in the last bar. The segmental SNR of normal coefficients with two more bits is also shown in Fig. 4 for a reference. Subsequently, the minimal number of additional bits that enable the modified coefficients to secure the equivalent segmental SNR to the normal coefficients is computed and summarized as shown in Table 1. For example, regarding 2 kbps harmonic codec and 16 kbps MDCT quantization, one more bit per frame allows the modified coefficients to have the same quantization performance as the normal coefficients.

The SNR analysis shows that in the worst case, two more bits are required for the equivalent quantization performance. However, this phenomenon never implies that the proposed look-ahead sharing degrades the quality of overall codec, because the amount of bit increase is small and the MDCT quantization only affects the enhancement layer. In this letter, this argument is verified indirectly by measuring the quality difference between the normal G.729.1 and the modified G.729.1 with two more bits for MDCT quantization. If two codecs have equivalent overall quality, then it is confirmed that two-bit-difference in MDCT quantization causes no effect on the quality of overall codec.

The test material consists of twelve 4 sec-long speech sentences spoken by six male and six female speakers. Eight listeners participate in the evaluation. For each sentence, the participants compare the quality of signal processed by the normal G.729.1 with that by the modified G.729.1 in terms of $-3 \sim +3$ comparison scale, where 0 corresponds to no perceptual difference. Figure 5 summarizes the evaluation results with 95% confidence interval, showing that two codecs provide the statistically equivalent perceptual quality. Consequently, the proposed look-ahead sharing for delay reduction causes no degradation in the quality of overall codec.

2.4 Encoder and Decoder

Figure 6 shows the encoder and decoder equipped with the

proposed MDCT structure. The window is assumed to be constant over the frames and any symmetric window of length $2N$ satisfying $w^2[n] + w^2[N - 1 - n] = 1$, such as sine and KBD (Kaiser-Bessel derived) windows, can be used [7]. In the encoder, the residual signal of the core layer is computed without overlap-and-add operation, and the incorrect MDCT operations caused by look-ahead sharing are included in the encoding. In the decoder, the output of the core layer is first obtained, and the error components are determined. Then, the IMDCT output is computed and the error components are eliminated through signal subtraction, resulting in the correct codec output. Finally, it is summed up with the output of the core layer, and the overall output is obtained. Additional operations at the proposed decoder include the computation of error components, $[-B_R w_3 + D_R w_1][-A_R w_4 + C_R w_2]$, and signal subtraction, which require $2N$ real multiplications and $2N$ real additions per frame.

By using the proposed structure, the overall algorithmic delay of the encoder and the decoder is reduced from $3N$ to $2N$. The frame size of conventional scalable speech codec such as G.729.1 is $N = 320$ with 16 kHz sampling rate, that corresponds to 20 msec. Assuming this frame size, for example, the proposed structure yields a delay reduction from 60 msec to 40 msec.

3. Conclusions

In this letter, the delay accumulation problem of the scalable codec with the cascaded DFT and MDCT was investigated. Such accumulation can be avoided by forcing two transforms to share a common look-ahead region, but incorrect operations in MDCT/IMDCT can occur due to look-ahead sharing. In this letter, the error components resulting from the above structure were determined by analyzing the MDCT/IMDCT operations, and it was confirmed that the error components are described in a closed form and are available in the decoder after the core layer. Hence, the error components need not be transmitted from the encoder to the decoder; rather, they can be computed and removed in the decoder to obtain the final correct results. The proposed structure reduces the codec delay by the frame size with an equivalent coding efficiency.

Acknowledgement

This work was supported by the Research Grant from Samsung Advanced Institute of Technology, Korea.

References

[1] A. McCree, "A 14 kbps wideband speech coder with a parametric highband model," Proc. IEEE ICASSP, pp.1153–1156, 2000.
[2] K.T. Kim, S.K. Jung, Y.C. Park, and D.H. Youn, "A new bandwidth scalable wideband speech/audio coder," Proc. IEEE ICASSP, pp.657–660, 2002.
[3] K. Koishida, V. Cuperman, and A. Gersho, "A 16-kbit/s bandwidth scalable audio coder band on the G.729 standard," Proc. IEEE ICASSP, pp.1149–1152, 2002.
[4] ITU G.729.1, "G.729 based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729," 2006.
[5] D.W. Griffin and J.S. Lim, "Multi-band excitation vocoder," IEEE Trans. Acoust. Speech Signal Process., vol.36, no.8, pp.1223–1235, 1988.
[6] ISO/IEC 14496-3, "Coding of Audio-Visual Objects—Part 3: Audio, Subpart 2, Speech Coding—HVXC," 1999.
[7] ISO/IEC 14496-3, "Coding of Audio-Visual Objects—Part 3: Audio, Subpart 4: General Audio Coding—AAC, TwinVQ, BSAC," 2001.
[8] J.P. Princen and A.B. Bradley, "Analysis/synthesis filter bank design based on the domain aliasing cancellation," IEEE Trans. Acoust. Speech Signal Process., vol.34, no.5, pp.1153–1161, 1986.