Transform Coding Based on Source Filter Model in the MDCT Domain

Jongmo Sung and Yun-Ho Ko

State-of-the-art voice codecs have been developed to extend the input bandwidth to enhance quality while maintaining interoperability with a legacy codec. Most of them employ a modified discrete cosine transform (MDCT) for coding their extended band. We propose a source filter model-based coding algorithm of MDCT spectral coefficients, apply it to the ITU-T G.711.1 super wideband (SWB) extension codec, and subjectively test it to validate the model. A subjective test shows a better quality over the standardized SWB codec.

Keywords: Source filter model, transform coding, MDCT, G.711.1 Annex D.

I. Introduction

Recently, many speech and audio codecs were developed by various standardization bodies, such as ITU-T, 3GPP, and MPEG. One of the most prevailing trends is to enhance the quality of a codec by extending it to a wider signal bandwidth while providing interoperability with the legacy codecs [1]-[3]. Another trend is to provide a consistent high quality for speech, music, and mixed contents over a broad range of bitrates [4]. The combination of time-domain coding and transform coding in an embedded or switched manner has been widely used to meet these trends. In the embedded codecs, the input signal is usually split into lower and higher band signals. The lower band signal is encoded with a core codec generating an interoperable bitstream, and the higher band signal is transformed into a modified discrete cosine transform (MDCT) domain and encoded with a dedicated transform coding algorithm accompanying a bandwidth extension [1]-[3].

Because the transform coding schemes used in most codecs directly encode the MDCT spectral coefficients by exploiting such input signal characteristics as the tonality, harmonicity, and stationarity, these approaches are somewhat heuristic and complicated. If we can establish a generic model that is less dependent on the input characteristics for the coding of an MDCT spectrum, a well-structured coding will be possible.

In this letter, we propose a source filter model for the coding of an MDCT spectrum and give a general description of the transform coding algorithm based on this model. To validate the model, we apply the proposed algorithm to the higher band coding of the ITU-T G.711.1 super wideband (SWB) extension codec [3] and conduct a subjective listening test.

II. Source Filter Model-Based Transform Coding

In the source filter model of speech production [5], the excitation of the filter is modeled as either an impulse train for voiced speech or as random noise for unvoiced speech. A time-varying digital filter represents the vocal tract and radiation characteristic. Similarly, as an MDCT spectrum contains both tonal and non-tonal components, the excitation in the source filter model for MDCT coding can be represented in a combination of impulses for the tonal component and noises for the non-tonal component. A block diagram of the proposed source filter model of an MDCT spectrum is shown in Fig. 1. Regarding the selection of the source filter type used in this model, a linear prediction (LP) filter can be a good candidate, as the prediction in the spectral domain mitigates the pre-echo artifacts that inevitably occur during transform coding [6].

Exploiting the concept of a source model for speech production, we can apply the efficient analysis methods used in

Manuscript received Aug. 29, 2012; revised Nov. 28, 2012; accepted Dec. 6, 2012.
This work was supported by the Broadcasting Technology R&D program of KCC/KCA (No.11921-02001).
Jongmo Sung (phone: +82 42 860 1243, jmseong@etri.re.kr) is with the Broadcasting & Telecommunications Media Research Laboratory, ETRI, Daejeon, Rep. of Korea.
Yun-Ho Ko (koyh@cnu.ac.kr) is with the Department of Mechatronics Engineering, Chungnam National University, Daejeon, Rep. of Korea.
http://dx.doi.org/10.4218/etrij.13.0212.0368

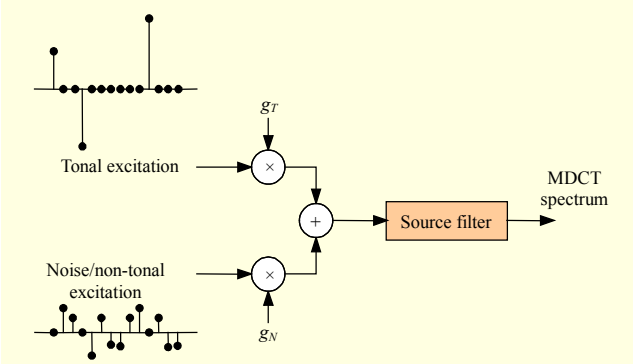

Fig. 1. Block diagram of source filter model of MDCT spectrum.

code-excited LP (CELP) [5] to the proposed model based on the analysis-by-synthesis scheme.

An input signal in the time domain is transformed into an MDCT spectrum. A set of coefficients of an LP filter is computed using an LP analysis on the MDCT spectrum. From the LP coefficients, an LP synthesis filter used for a codebook search is formed to compute the synthesized spectrum. Tonal codebook and non-tonal codebook searches are performed to determine the excitation of the LP filter.

For a tonal codebook, we find the best combination of tonal components to minimize the difference between the synthetically generated spectrum and the original spectrum. Then, the tonal contribution is subtracted from the original spectrum to find the target spectrum for the non-tonal codebook. The non-tonal codebook is determined to minimize the difference between the LP synthesized spectrum and the target spectrum. Using the analysis-by-synthesis coding scheme, we can mimic the adaptive codebook and fixed codebook search methods of a CELP-type codec to find the best tonal codebook and the best non-tonal codebook.

III. G.711.1-SWB Implementation Using Source Filter Model-Based Algorithm

To validate the proposed model-based transform coding algorithm, we apply the algorithm to G.711.1 Annex D, which provides a scalable SWB speech and audio coding algorithm operating from 96 kb/s to 128 kb/s, depending on the G.711.1 core mode. Specifically, we replace the first SWB layer of 16 kb/s in the G.711.1 Annex D with a layer adopting the model-based transform coding algorithm. Figure 2 shows the overall encoding algorithm of the proposed G.711.1-SWB codec.

A 32-kHz sampled input signal is divided into two 16-kHz sampled lower band and higher band signals using a quadrature mirror filter (QMF). A lower band spectrum spanning from 0 Hz to 7 kHz is generated by concatenating the 0-Hz to 4-kHz band from a narrowband MDCT and the 4-kHz to 7-kHz band recovered from the G.711.1 bitstream. The target higher band spectrum for a codebook search comprises the 7-kHz to 8-kHz band from the G.711.1 core encoder and the 8-kHz to 14-kHz band from the higher band MDCT.

First, the higher band spectrum is normalized using a quantized global gain, computed as follows:

$$g_{\text{glob}} = \text{round}\left(\log_2 \left(\sqrt{\frac{1}{70}\sum_{k=0}^{69} X_{\text{HB}}(k)^2} + \varepsilon_{\text{rms}} \right) \right), \quad (1)$$

where $\varepsilon_{\text{rms}}=2^{-16}$. An LP analysis is performed on the normalized higher band spectrum to compute the sixth-order LP coefficients. The LP coefficients are transformed into the line spectral frequencies and are vector-quantized. An LP residual spectrum, $R_{\text{HB}}(k)$, is computed as follows:

$$R_{\text{HB}}(k) = \bar{X}_{\text{HB}}(k) + \sum_{i=1}^{6} \hat{a}_i \bar{X}_{\text{HB}}(k-i), k = 0,...,69, \quad (2)$$

where $\hat{a}_i$, i=1,…,6, are the LP coefficients and $\bar{X}_{\text{HB}}(k)$ is a higher band spectrum normalized using the quantized global gain. The tonal codebook is based on a structure using an interleaved single pulse permutation design. In the codebook structure given in Table 1, each codebook vector contains eight non-zero pulses. Each pulse is represented in its sign, amplitude, and position. The tonal codebook vector is constructed by taking a zero vector with a dimension of 70 and putting the eight non-zero pulses at the found location multiplied by their corresponding sign and amplitude:

$$\begin{aligned} T_i(k) &= s_i \delta(k - m_i),\ i = 0,...,7, k = 0,...,69, \\ T(k) &= \sum_{i=0}^{7} g_T(i) T_i(k),\ k = 0,...,69, \end{aligned} \quad (3)$$

where $\delta(k)$ is a unit pulse. To reduce the computational complexity of a codebook search, the initial tonal codebook vector is obtained by searching the pulse with the largest amplitude in each track. Each of the eight pulses in the initial codebook vector is sequentially replaced by another pulse on the same track, and the mean squared error (MSE) between the original spectrum and LP synthesized spectrum for the updated tonal codebook vector is computed. The best pulse on the track is searched by minimizing (4).

$$E = \sum_{k=0}^{69} \{X(k) - Y(k)\}^2 = \sum_{k=0}^{69} \left\{ X(k) - \left[T(k) - \sum_{i=1}^{6} \hat{a}_i Y(k-i) \right] \right\}^2, \quad (4)$$

where $Y(k)$ is the LP-filtered tonal codebook vector for each pulse combination. This procedure is repeated for every track to produce the best tonal codebook vector.

In the non-tonal codebook, we implement a filter similar to a pitch filter, which has delay and gain parameters. The target spectrum used for the non-tonal codebook search is obtained by subtracting the tonal codebook contribution from the

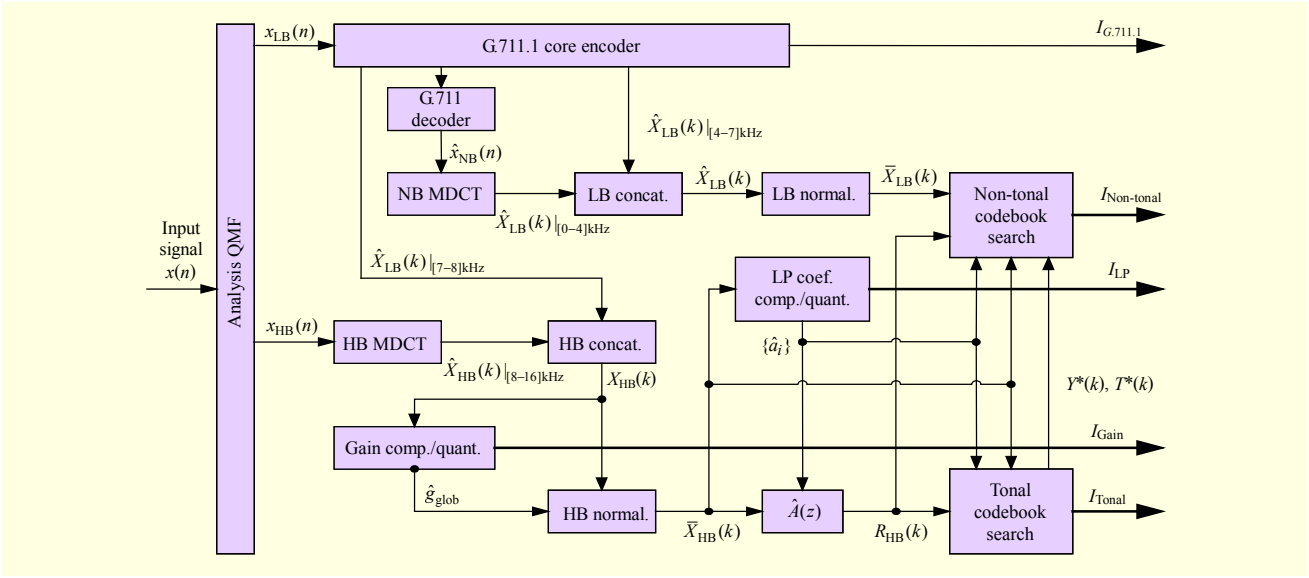

Fig. 2. Block diagram of proposed G.711.1-SWB encoder.

Table 1. Structure of tonal codebook.

Pulse	Sign	Amplitude	Position
i_0	s_0: ±1	$g_7(0)$	m_0: 0, 8, 16, 24, 32, 40, 48, 56
i_1	s_1: ±1	$g_7(1)$	m_1: 1, 9, 17, 25, 33, 41, 49, 57
i_2	s_2: ±1	$g_7(2)$	m_2: 2, 10, 18, 26, 34, 42, 50, 58
i_3	s_3: ±1	$g_7(3)$	m_3: 3, 11, 19, 27, 35, 43, 51, 59
i_4	s_4: ±1	$g_7(4)$	m_4: 4, 12, 20, 28, 36, 44, 52, 60
i_5	s_5: ±1	$g_7(5)$	m_5: 5, 13, 21, 29, 37, 45, 53, 61
i_6	s_6: ±1	$g_7(6)$	m_6: 6, 14, 22, 30, 38, 46, 54, 62
i_7	s_7: ±1	$g_7(7)$	m_7: 7, 15, 23, 31, 39, 47, 55, 63

normalized higher band spectrum as follows:

$$X'_{\mathrm{HB}}(k) = \bar{X}_{\mathrm{HB}}(k) - Y^{*}(k), k = 0,...,69\,, \tag{5}$$

where $Y^{*}(k)$ is the LP-filtered contribution of the best tonal codebook, $T^{*}(k)$. The non-tonal codebook search basically utilizes the property in which the higher band spectrum is correlated with the lower band spectrum. The normalization process for the lower band spectrum of 0 Hz to 7 kHz is done before the non-tonal codebook search. To reduce the complexity of the non-tonal codebook lag search, we perform an open-loop search and then a closed-loop search in the search range around the candidate lag obtained in the open-loop search (Table 2(a)).

An open-loop lag is computed based on a normalized cross-correlation between the normalized lower band spectrum, $\bar{X}_{\mathrm{LB}}(k)$, and LP residual spectrum subtracting the best tonal codebook, $R_N(k)=R_{\mathrm{HB}}(k)-T^{*}(k)$.

Table 2. Subband boundaries, number of coefficients, and search range for non-tonal codebook.

	j	$b_{\mathrm{op}}(j)$	$N_{\mathrm{op}}(j)$	$L_{\mathrm{op}}(j)$	$U_{\mathrm{op}}(j)$
(a) Open-loop	0	0	35	0	32
	1	35	35	0	32
	2	70	-	-	-
	j	$b_{\mathrm{cl}}(j)$	$N_{\mathrm{cl}}(j)$	$L_{\mathrm{cl}}(j)$	$U_{\mathrm{cl}}(j)$
(b) Closed-loop	0	0	18	$0 \le Lag_{\mathrm{op}}(0)-8$	$Lag_{\mathrm{op}}(0)+8 \le 32$
	1	18	17	$Lag_{\mathrm{cl}}(0)+N_{\mathrm{cl}}(0)-8$	$Lag_{\mathrm{cl}}(0)+N_{\mathrm{cl}}(0)+8$
	2	35	18	$0 \le Lag_{\mathrm{op}}(1)-8$	$Lag_{\mathrm{op}}(1)+8 \le 32$
	3	53	17	$Lag_{\mathrm{cl}}(2)+N_{\mathrm{cl}}(2)-4$	$Lag_{\mathrm{cl}}(2)+N_{\mathrm{cl}}(2)+4$
	4	70	-	-	-

$$Lag_{\mathrm{op}}(j) = \underset{L_{\mathrm{op}}(j) \le l < U_{\mathrm{op}}(j)}{\arg\max} \left\{ \frac{\sum_{k=b_{\mathrm{op}}(j)}^{b_{\mathrm{op}}(j+1)-1} R_N(k)\bar{X}_{\mathrm{LB}}(k+l)}{\sqrt{\sum_{k=b_{\mathrm{op}}(j)}^{b_{\mathrm{op}}(j+1)-1} \bar{X}_{\mathrm{LB}}(k+l)\bar{X}_{\mathrm{LB}}(k+l)}} \right\}, j=0,1. \tag{6}$$

A closed-loop search of the non-tonal codebook then minimizes the MSE between the target spectrum and the LP synthesized spectrum for each subband specified in Table 2(b). This is achieved by maximizing the following term:

$$\sum_{k=b_{\mathrm{cl}}(j)}^{b_{\mathrm{cl}}(j+1)-1} X'_{\mathrm{HB}}(k)Z_j(k+l) \Big/ \sqrt{\sum_{k=b_{\mathrm{cl}}(j)}^{b_{\mathrm{cl}}(j+1)-1} Z_j(k+l)Z_j(k+l)}, \tag{7}$$
$$L_{\mathrm{cl}}(j) \le l < U_{\mathrm{cl}}(j),\ \ j = 0,1,2,3,$$

where $Z_j(k+l)$ is the spectrum synthesized by an excitation of the normalized lower band spectrum of the j-th subband at delay l.

The overall number of bits of the model-based transform coding layer is 80 bits per frame. The global gain and LP coefficients are encoded at 5 bits and 6 bits, respectively. The 42 bits for the tonal codebook and 27 bits for the non-tonal codebook are used. The bit allocation is shown in Table 3.

Meanwhile, the worst-case computational complexities of an encoder and decoder based on an ITU-T STL2009 [7] are 28.4 and 3.6 weighted million operations per second, respectively.

We carry out subjective tests to assess the quality of the proposed G.711.1-SWB codec using a triple stimulus/hidden reference/double blind method with a five grade impairment scale ranging from 1.0 (very annoying) to 5.0 (imperceptible). The test method is compliant with ITU-R BS.1116-1 [8]. The codec is tested at the operating mode of R3sm, that is, 96-kb/s SWB mono with a G.711.1 80-kb/s core. Eight expert listeners participate in the test, which is performed for two categories, that is, speech and music/mixed contents. An overview of the subjective listening test is given in Table 4.

Figure 3 shows a summary of the subjective test results. As shown in the figure, for all cases, the quality provided by the proposed codec is better than or comparable to the quality provided by G.711.1 Annex D. Noticeably, the proposed codec outperforms the reference codec for speech at –26 dBov.

Table 3. Bit allocation of model-based transform coding layer.

Parameter		No. of bits	Total
Global gain		5 bits	5 bits
LP coefficients		6 bits	6 bits
Tonal codebook	Positions	3 bits/pulse × 8 pulses	24 bits
	Signs	1 bit/pulse × 8 pulses	8 bits
	Amplitudes	5 bits/vector × 2 vectors	10 bits
Non-tonal codebook	Lags	(5+4+5+3) bits	17 bits
	Gains	5 bits/vector × 2 vectors	10 bits
Total			80 bits

Table 4. Overview of subjective listening test.

Category	Speech	Music and mixed contents
Input level	–16/–26/–36 dBov	–26 dBov
Test materials	3 female and 3 male talkers 4 samples/talker	2 types of music and 2 types of mixed contents 4 samples/type
No. of subjects	8	8
Ref. codec	G.711.1 Annex D @ R3sm (96 kb/s)	

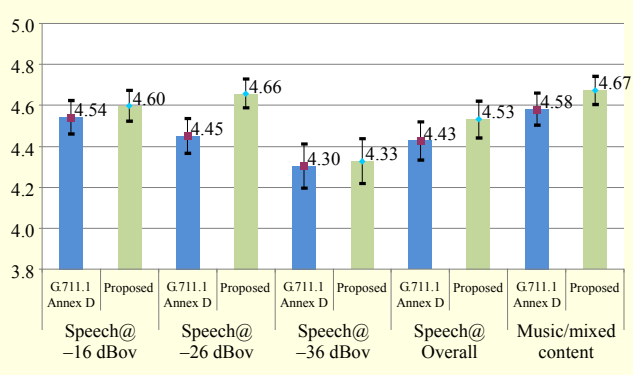

Fig. 3. Subjective test results.

IV. Conclusion

Distinctly different from conventional transform coding schemes using the direct quantization of spectral coefficients, we proposed a source filter model-based transform coding algorithm for an MDCT spectrum. The model is composed of an LP filter and an excitation of the tonal and non-tonal codebooks having a similar structure to a conventional CELP codec. We realized the model in the higher band coding of the G.711.1-SWB codec and evaluated the quality of the codec subjectively to validate the model. Compared with G.711.1 Annex D, the proposed codec shows desirable quality for both speech and music/mixed contents.

References

[1] ITU-T Rec. G.729.1, *An 8-32 kbit/s Scalable Wideband Coder Bistream Interoperable with G.729*, 2006.

[2] ITU-T Rec. G.718 Annex B, *Superwideband Scalable Extension for G.718*, 2008.

[3] ITU-T Rec. G.711.1 Annex D, *Wideband Embedded Extension for G.711 PCM: New Annex D with Superwideband Extension*, 2010.

[4] T. Lee et al., "Adaptive TCX Windowing Technology for Unified Structure MPEG-D USAC," *ETRI J.*, vol. 34, no. 3, June 2012, pp. 474-477.

[5] A.M. Kondoz, *Digital Speech*, Chichester, UK: John Wiley & Sons, 1994.

[6] J. Herre, "Temporal Noise Shaping, Quantization and Coding Methods in Perceptual Audio Coding: A Tutorial Introduction," *17th Int. AES Conf.*, 1999, pp. 312-325.

[7] ITU-T Rec. G.191, *Software Tools for Speech and Audio Coding Standardization*, 2010.

[8] ITU-R Rec. BS.1116-1, *Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems*, 1997.