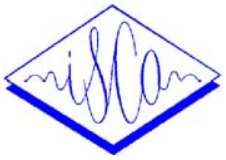
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Temporal Normalization Techniques for Transform-type Speech Coding and Application to Split-band Wideband Coders

Kyung-Tae Kim, Sung-Kyo Jung, MiSuk Lee†, Hong-Goo Kang, and Dae Hee Youn

Department of Electrical and Electronic Eng., Yonsei University, Korea
† Broadband Convergebce Network Research Division, ETRI

10.21437/Interspeech.2004-707

Abstract

In this paper we present an efficient coding method for the upper band(4-7kHz) of wideband(0.5-7kHz) speech coding based on a band-split approach. Due to the impulse-like characteristics in upper band signal, it is very difficult to efficiently quantize the signal at low bit-rate when we use transform coding techniques. We propose two temporal normalization techniques, *direct temporal energy normalization* and *frequency domain linear prediction*, to reduce the extremely noticeable artifacts. Simulation results show that the proposed algorithm successfully encodes the upper band signal, and the new split-band type wideband coder adopting the proposed technology provides better quality than 56 kbit/s ITU-T G.722 at the bit-rate of 20 kbit/s.

1. INTRODUCTION

The recent speech coding standards provide high speech quality that is sufficient for deploying them to the commercial market. Due to fast growing trend of using packet communication systems, future speech coding standards should be able to adopt various system constraints as well as to provide audio bandwidth scalability. Recently, ITU identified the importance of these application areas and defined the terms of reference(ToR) for new variable bit-rate(VBR) codecs [1]: multi-rate source controlled VBR(MSC-VBR) approach and embedded VBR(EV) approach.

Several scalable coding schemes were proposed to support both bandwidth scalability and interoperability with the conventional standard coders such as G.723.1 as well as G.729 [2][3]. In [4], Jung et al adopts a split-band approach, where the input signal is decomposed into two equal frequency bands by means of quadrature mirror filters (QMF) [5]. The lower band speech is coded with G.723.1. To further improve the perceptual quality of lower band signals, they employ an additional coding unit. The upperband signal is encoded using transform coding based on modified discrete cosine transform (MDCT) [6]. The coder at a bit-rate of 19.4 kbit/s provides comparable speech quality to the ITU-T 24 kbit/s G.722.1 coder [7] while it also has interoperability with G.723.1. However, the approach presented some detectable artifacts at the onset regions. Moreover, slight amounts of harmonic noise can be detected all over the decoded speech, especially for male speech. All the problems are caused by lack of bit availability and pre-echo artifacts while a transient signal is being coded in a spectral representation domain.

This paper focuses on improving the coding performance of the upper band signal. In order to remove pre-echo artifacts we try to flatten temporal energy distribution because the artifacts occur when the distribution of signal energy varies severely such as at the onset or in pitched signal regions. We try two approaches for flattening the temporal energy distribution: direct temporal energy normalization and frequency domain linear prediction. Two methods are utilized to make a wideband coder interoperable with ITU-T G.729 at the bit-rate of 20 kbit/s. As a result, the proposed wideband coder with temporal normalization provides better speech quality than the 56 kbit/s ITU-T G.722 [8] coder.

2. GENERAL SPLIT-BAND WIDEBAND CODER

A block diagram of the wideband coder with split-band approach is shown in Figure 1. The coder splits the input speech into lower band (0-4 kHz) and upper band (4-8 kHz) by means of QMF. For the core layer, encoding with all kinds of narrow band coders is applicable. The enhancement layer consists of two layers, the lower band enhancement layer and the upper band coding layer. The lower band enhancement layer plays a role in improving the quality of the core layer in the 0-4 kHz band. The upper band coding layer handles the signal of the 4-7 kHz band with various coding methods, in which time domain and frequency domain coding methods can be used. In the time domain coding approach, the upper band signal is quantized with PCM or CELP-type coders, while in frequency domain coding the signal is first transformed to the frequency domain with FFT, DCT or MLT and then the transformed coefficients are quantized. When the up-

This work was supported by the Electronics and Telecommunications Research Institute (ETRI).

per band signal is coded at a low bit rate, the quantization noise from the time domain approach sounds, in general, more offensive to the ear than the one from the frequency domain.

The wideband coder presented in [4] showed good quality and interoperability with the core coder, but it still had the inveterate problem due to a frequency domain coding method, so-called pre-echo problem. The reason for pre-echo artifacts is known to be the inappropriate temporal spread of quantization noise. These artifacts occur when a transient signal or a pitched signal is coded in a spectral representation manner. In this case, the quantization noise is spread out over the entire window length of the filterbank but it is not masked by the signal itself. While an upper band speech signal is characterized as noise-like, it also has a periodic energy distribution and "attack" in some regions. We are able to observe weak whisper-like noise from the wideband coder [4]. The noise exists all around the coded speech, especially male speech with a low fundamental frequency.

3. TEMPORAL ENERGY ENVELOP MODELING

In order to avoid a pre-echo problem, care has to be taken to maintain the appropriate temporal characteristics of the target signal. We tried to model the temporal characteristics of the upper band signal: direct temporal energy normalization and frequency domain linear prediction.

3.1. Direct temporal energy normalization

Figure 2 shows a block diagram of the upper band coder with direct temporal energy normalization. In this approach, the decimated upper band signal is first normalized in the time domain and transformed into spectral domain components by MDCT. The MDCT coefficients are grouped into subband, the separated by gains(scale factor) and shapes(normalized coefficients), which are quantized and transmitted to the decoder together with the temporal energy.

Normalization processing is performed for each segment being a group of samples in the time domain. The

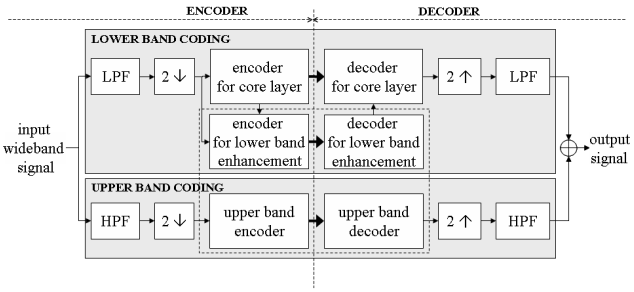


Figure 1: Conventional split-band wideband coder.

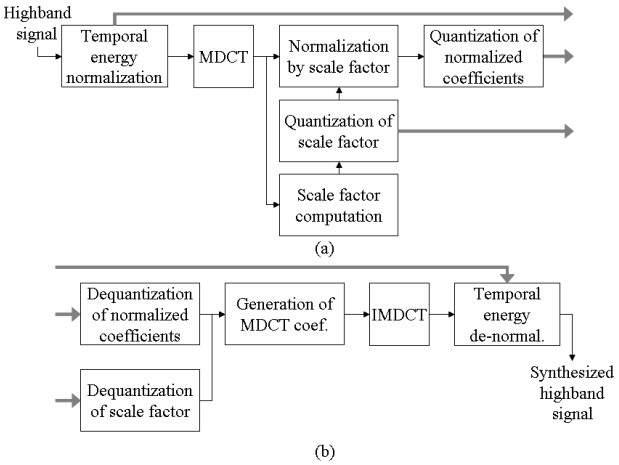


Figure 2: Direct temporal energy normalization approach. (a) Encoder. (b) Decoder

smaller the segment size for temporal normalization, the larger the amount of bits required to quantize the temporal energy. However, a large segment size can not achieve a favorable result from the temporal normalization. With a segment of 10 samples(1.25 msec) we can achieve proper normalization. The segment length of 10 samples is a much shorter period of time than pre- and post-temporal masking. We can, therefore, remove the pre-echo problem without restraint. The ith segmental energy, $\varepsilon_x(i)$ of the frame of 10 msec becomes

$$\varepsilon_x(i) = \sqrt{\frac{1}{10}\sum_{n=10i}^{10i+9} x_H^2(n)}, \quad 0 \le i < 8 \qquad (1)$$

where $x_H(n)$ represents the decimated upper band signal. A temporal normalized signal, $x_n(n)$ is found as

$$x_n(n) = \frac{x_H(n)}{\varepsilon_x(\lfloor n/10 \rfloor)}, \quad 0 \le n < 80 \qquad (2)$$

As shown in Figure 3, the segmental energy is quasi-periodic, while the waveform of the upper band signal is not. Utilizing the characteristics of the temporal energy contour, we can quantize the block energy with a long-term predictor. The lag of the long-term predictor is set to the same value as the pitch lag, which is obtained from the core coder. The reason for this is because a period of the upper band signal is highly correlated with one of the lower band signal.

3.2. Frequency domain linear prediction

Linear prediction in the frequency domain can be used as an another method for modeling the temporal characteristics of the upper band signal. It is well-known that signals with anything but flat spectrum can be coded efficiently either by directly coding spectral values, or by

applying predictive coding methods to the time signal. Considering time-frequency duality, efficient coding of transient signals can thus be achieved through employing predictive coding methods to the spectral data by carrying out a prediction across the frequency domain [10]. Figure 4 shows the block diagram of an upper band coder with the linear prediction in the frequency domain. An MDCT is preceded by a temporal normalization procedure contrary to the case of direct temporal energy normalization. Linear prediction is performed in the MDCT domain. Residual MDCT coefficients obtained from the linear prediction analysis are decomposed into scale factors and shapes, and then quantized and transmitted to the decoder as in the direct temporal energy normalization method.

The decision of the prediction order relates to the number of peaks in temporal energy contour, which is closely related to the pitch period. Considering human pitch range, a maximum of 4 pulses are possible in a frame of 10 msec. To represent 4 pulses, the order of linear predictor is set to 8. Although the prediction order is fixed in this paper, it can adaptively be adjusted depending on the pitch period, of which method requires less bits than the fixed one.

In order to quantize the linear prediction coefficients, we convert them to line spectrum pairs (LSPs). Since there is no interframe correlation, interframe predictor that enhances performance of quantizer is not available. However, a more effective LSP quantizer may be introduced. First, the energy contour of the upper band signal is quasi-periodic and its period is close to the pitch lag of the core coder. In addition, the positions of the pitch pulses in both lower band signal and upper band signal are similar to each other. These parameters from the core layer might help to quantize the LSP for the upper band coding. Since the goal of this paper is to validate the possibility of frequency domain linear prediction, we quantize the LSP while disregarding the parametric information from the core coder.

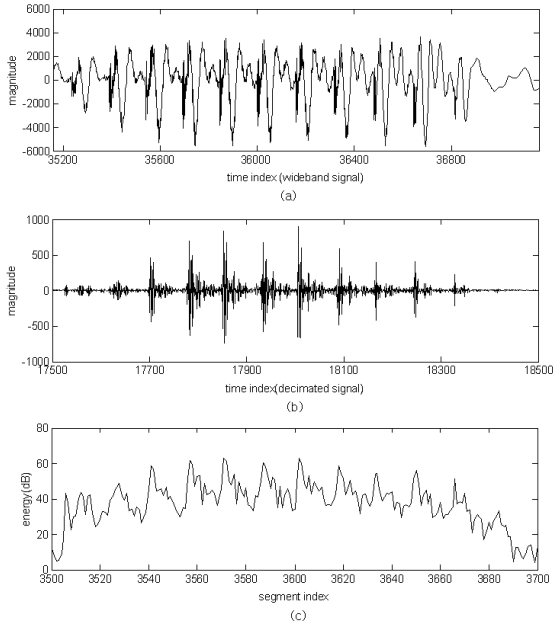


Figure 3: Comparison of temporal energy contour and speech signals. (a) wideband speech. (b) decimated upperband speech signal. (c) temporal energy contour.

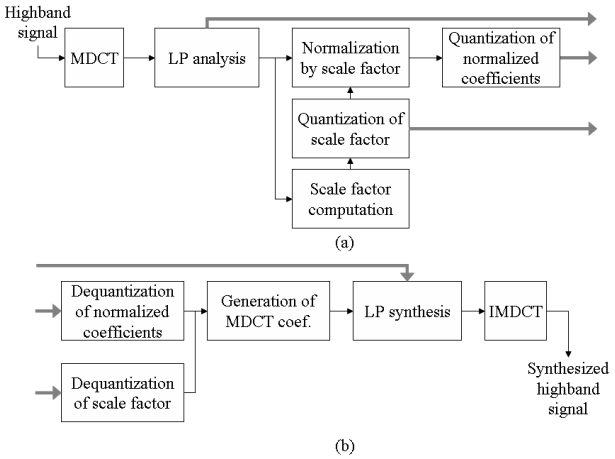


Figure 4: Frequency domain linear prediction approach. (a) Encoder. (b) Decoder

Table 1: Bit allocation for the upper band coder with direct temporal energy normalization

parameters	**bits 4kHz← →7kHz**
temporal normalization	8seg×4bits
band energies (differential coding)	(3,3,2,2,2,1)
shape vector (weighted interleaving VQ)	35
total	80

4. IMPLEMENTATION AND PERFORMANCE EVALUATION

Bit allocation tables for two approaches are given in Table 1 and Table 2. Scale factors for 6 bands are differential-coded in intra-frame, and normalized MDCT coefficients are coded by a weighted interleaving vector quantizer[11], which employs adaptive weighted matching criteria instead of conventional adaptive bit allocation. We have achieved the optimal bit allocation through observing and comparing the sensitivity of each param-

Table 2: Bit allocation for the upper band coder with frequency domain linear prediction

parameters	**bits 4kHz← →7kHz**
LSP	8
band energies (differential coding)	(4,3,3,2,2,2)
shape vector (weighted interleaving VQ)	56
total	80

Table 3: SegSNR for two temporal energy normalization methods (for 96 Korean samples in NTT multilingual database)

	direct temporal energy normalization	**frequency domain linear prediction**
segSNR	4.1703 dB	6.5176 dB

eter by varying the number of the allocated bits. Direct temporal energy normalization requires many more bits for temporal normalization than frequency domain linear prediction. As shown in Table 3, the frequency domain linear prediction approach presents a higher segSNR than direct temporal normalization does. However, the two methods show better segSNR scores than that of the previous work, [4], whose upperband coder gives a segSNR of 3.9679.

Table 4: Result of MOS test (for 4 female and 4 male Korean samples in NTT multilingual database)

	Original Speech	**G.722 56 kbit/s**	**Temp. norm.1** [a]	**Temp. norm.2** [b]
Female	4.71	4.08	4.12	4.26
Male	4.50	3.66	3.92	4.19

[a] direct temporal energy normalization
[b] frequency domain linear prediction

We have designed a wideband coder using the temporal energy envelop model, which operates at 20 kbit/s. The 40 bits per frame are allocated for lowerband enhancement, and all the bits are used in an additional fixed codebook [9]. A wideband coder with the frequency domain linear prediction approach produces a more natural sound, while a wideband coder with the direct temporal energy normalization produces some quantization noise similar to that detected in CELP-type coders. To compare perceptual quality differences, we performed MOS tests involving 20 listeners. Table 4 shows the results of the tests. The results imply that the frequency domain linear prediction is a more effective method for the upper band coding of wideband speech coders.

5. CONCLUSION

The primary objective of this paper is to enhance the quality of the wideband speech coder that we had previously proposed in [4]. In this paper, we proposed two methods, direct temporal energy normalization and frequency domain linear prediction to remove pre-echo artifact that caused a critical problem in the previous wideband coder. Consequently, wideband coders that apply two upperband coders offer better quality than 56 kbit/s G.722. It is open to further discussion how to smooth out the temporal energy contour without having artifacts in consecutive segments when using the direct temporal energy normalization approach. To find the optimal LPC order for each frame and to better quantize the LSP information for the frequency domain linear prediction approach is another issue.

6. References

[1] ITU-T SG 16 Tdoc.25, "Report of Q.24/16 previous meeting (Geneva, Nov. 2000)," Jun. 2001.

[2] ITU-T Rec. G.723.1, "Dual-rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s," May 1996.

[3] ITU-T Rec. G.729, "Coding of speech at 8 kbit/s using conjugate-structure algebraic code-excited linear prediction (CS-ACELP)," May 1996.

[4] S.-K. Jung, K.-T. Kim, and H.-G. Kang, "A bit-rate/bandwidth scalable speech coder based on ITU-T G.723.1 standard," in *Proc. ICASSP*, 2004, to be published.

[5] P.P.Vaidyanathan, *Multirate Systems and Filter Bank*, Prentice Hallk 1993

[6] J.P.Princen and A.Bradley, "Analysis/synthesis filter bank design based on time-domain aliasing cancellation," in *IEEE Trans. on ASSP*, vol.34, pp.1153-1161, Oct.1986.

[7] ITU-T Rec. G.722.1, "Coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss," Sep.1999

[8] ITU-T Rec. G.722, "7 kHz Audio-coding within 64 kbit/s," Sep.1999.

[9] S.-K. Jung, K.-T. Kim, H.-G. Kang, and D.H.Youn, "A cascaded algebraic codebook structure to improve the performance of the speech coders," in *Proc. ICASSP*, pp.II.173-II.176, Apr. 2003

[10] Jürgen Herre and James D. Johnston, "Enhancing the Performance of Perceptual Aduio Coders by Using Temporal Noise Shaping(TNS)," *101st AES*, 1996, Preprint 4384

[11] Takehiro Moriya and Masaaki Honda, "Transform Coding of Speech Using a Weighted Vecotr Quantizer," *IEEE Journal of Selected Areas in Communications*, vol.6, no.2, Feb. 1988.