Dual-mode AVQ Coding Based on Spectral Masking and Sparseness Detection for ITU-T G.711.1/G.722 Super-wideband Extensions

Masahiro Fukui[1], Shigeaki Sasaki[1], Yusuke Hiwasaki[1], Kurihara Sachiko[1], Yoichi Haneda[1]

[1] NTT Cyber Space Laboratories, NTT Corporation, Japan

{fukui.masahiro, sasaki.shigeaki, hiwasaki.yusuke, kurihara.sachiko, haneda.yoichi}@lab.ntt.co.jp

Abstract

ITU-T Recommendations G.711.1 Annex D and G.722 Annex B, which are super-wideband (50–14,000 Hz) extensions to G.711.1 and G.722, have been recently standardized. This paper introduces a new coding method proposed and employed in the above ITU-T standards. The proposed coding method employs an adaptive spectral masking of the algebraic vector quantization (AVQ) for MDCT-domain non-sparse signals. The adaptive spectral masking is switched on and off based on MDCT-domain sparseness analysis. When the target MDCT coefficients are categorized as non-sparse, masking level of the target MDCT coefficients is adaptively controlled using spectral envelope information. The performance of the proposed method as a part of the ITU-T G.711.1 Annex D is evaluated in comparison with the ordinary AVQ. Subjective listening test results show that the proposed method improves the sound quality more than 0.1 points with a five grade scale in average of speech, music and mixed content, and the significance of the improvement is validated.

Index Terms: speech and audio coding, Standardization, ITU-T G.711.1 Annex D, ITU-T G.722 Annex B, super-wideband extension, algebraic vector quantization

1. Introduction

An increasing number of telecommunication systems (i.e., video conferencing, videophones, etc.) are being used over broadband networks. For better sound clarity and less listener fatigue, the frequency band of some systems is supported up to the near-CD-audio wideband, such as, the 14-kHz super-wideband (SWB) of a 32-kHz sampling frequency.

Speech coding is a key technology in telecommunication systems. It is necessary for speech transmission even if a broadband network is used. The recent codecs for SWB signals are generally designed for various sound sources (e.g., speech, music, and mixed content), unlike conventional narrow-band or wideband codecs, which handle speech only. The SWB codecs are required for high-definition telecommunication systems, e.g., telepresence systems. On the basis of the need for these SWB codecs, several standardization activities have been conducted recently.

A fast and efficient coding method is more important for processing the greater number of SWB samples. One of the well-known coding methods is the algebraic vector quantization (AVQ) [1], which is a sparse type of vector quantization [2]. Since the speech signals can be assumed to be sparse signals, the codevectors of the AVQ are represented with a few non-zero components. The AVQ is also applied to the frequency domain coding using modified discrete cosine transform (MDCT) because the sparseness can be found in not only speech but a part of music sound items in that domain. However, the lack of available bit budget occasionally causes the perceptual degradation of sound quality when quantizing "non-sparse" frequency-domain signals (e.g., noise component included in music).

To solve this problem, an adaptive spectral masking of AVQ for MDCT-domain non-sparse signals using a masking threshold is proposed. The masking threshold is adaptively calculated from the spectral envelope. A method for switching the new spectral masking on and off based on the frequency-domain sparseness analysis is also proposed. This method has been added to ITU-T Recommendations G.711.1 Annex D [3] and G.722 Annex B [4] (SWB extensions of G.711.1 [5] and G.722 [6]), which were standardized in November 2010.

The remainder of this paper is organized as follows. Section 2 describes the conventional AVQ. Section 3 presents the AVQ with new spectral masking and the sparseness detection method for switching the spectral masking on and off. Section 4 provides implementation details of our method in ITU-T G.711.1 Annex D and ITU-T G.722 Annex B. Section 5 describes the subjective evaluation results, and the paper is concluded in Section 6.

2. Ordinary AVQ

An encoding block diagram example of the ordinary AVQ in the frequency domain is shown in Figure 1. First, by MDCT, the input signal is transformed to MDCT coefficient $S(k)$, where $k = 0,\ldots,L\text{-}1$. L is the number of samples in the frame. The L MDCT coefficients are split into N sub-bands. Second, the spectral envelope $f_{\mathrm{rms}}(\lfloor k/N \rfloor)$ is computed as a set of root mean square (RMS) values per sub-band and quantized. Finally, the input MDCT coefficient $S(k)$ is normalized using the quantized spectral envelope $\hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor)$ as

$$S^{\mathrm{norm}}(k) = \frac{S(k)}{\hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor)}, \quad (1)$$

and the obtained normalized coefficient $S^{\mathrm{norm}}(k)$ is encoded using the AVQ.

3. Adaptive Spectral Masking and Sparseness Detection

The described conventional method assumes that the speech signals are sparse, and the normalized coefficients are

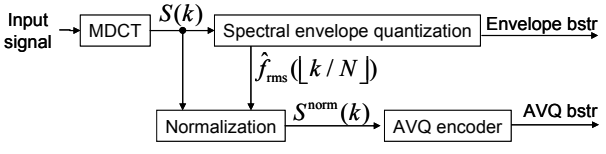

Figure 1: *Encoding block diagram of ordinary AVQ (bstr: bitstream).*

represented by a pulse-like (sparse) codevector composed of many zero and a few non-zero components. In contrast, the new method is applied to input MDCT coefficients characterized as "non-sparse".

The encoding block diagram of the new method is shown in Figure 2. In this figure, "sparse mode" is the ordinary AVQ, described in Section 2, and "non-sparse mode" is the new non-sparse type coding. Given the input MDCT coefficients $S(k)$, the mode selector decides the encoding mode according to the sparseness of the $S(k)$. If the sparse mode is selected, the AVQ encodes the normalized coefficients in the MDCT domain using the ordinary AVQ. Otherwise, only the coefficients, of which the energies are higher than the masking threshold, are encoded using the non-sparse mode,. The non-sparse mode and the mode selector are presented in detail in the following subsections.

3.1. Non-sparse mode

A conceptual diagram of non-sparse mode is shown in Figure 3. The sparse residual coefficients, which are obtained by subtracting the masking threshold from the input MDCT coefficient, are vector-quantized using the non-sparse mode. The masking threshold is adaptively calculated from the spectral envelope. The decoder reconstructs the non-sparse coefficients by adding the decoded residual coefficients to the offset based on the masking threshold or spectral envelope. In the non-sparse type of signals, zero sequences in the decoded coefficients could be easily detected as audible degradations. In this mode, the perceptual sound degradation is reduced by filling the components below the masking threshold with the noise.

3.1.1. Computing and encoding residual coefficients

The frequency components, of which the amplitudes among the input MDCT coefficients are higher than the masking threshold, are converted into the differences between their magnitude and the masking threshold, and the rest of components are set to zero. As a result, the obtained residual coefficients are composed of many zero and a few non-zero components. These non-zero components are dominant in the input coefficients. Since the residual coefficients are sparse compared to the input MDCT coefficients, they are effectively encoded using the sparse type of vector quantization, such as AVQ.

The residual coefficient $D(k)$ is computed from the input MDCT coefficient $S(k)$ as follows:

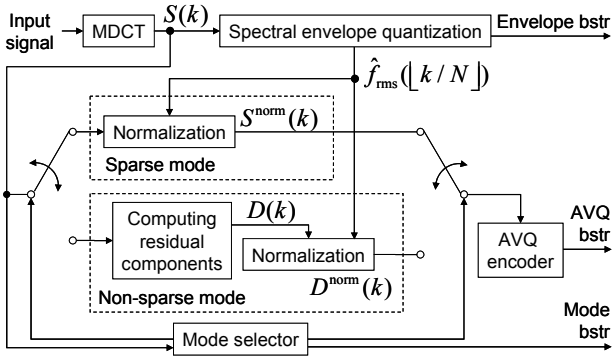

Figure 2: *Encoding block diagram of our method (bstr: bitstream).*

$$D(k) = \mathrm{sgn}\left(S(k)\right) \cdot \max\left(\left|S(k)\right| - \alpha \hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor), 0\right), \quad (2)$$

where $\mathrm{sgn}(\cdot)$ is the polarity of the MDCT coefficient, $\max(,)$ is the maximum value selection, and $\alpha \hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor)$ is the masking threshold used for calculating the sparse residual coefficients. α is the constant factor for adjusting the sparseness of the residual coefficients. The obtained error spectrum is normalized as

$$D^{\mathrm{norm}}(k) = \frac{D(k)}{\beta \hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor)}. \quad (3)$$

Because the RMS of $D(k)$ for normalization is not transmitted, $\beta \hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor)$ is used in place of the RMS of $D(k)$. β is the constant factor to close $\hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor)$ to the RMS value of $D(k)$, which is less than or equal to one. The normalized coefficient $D^{\mathrm{norm}}(k)$ is encoded using the AVQ and transmitted to the decoder.

3.1.2. Reconstructing non-sparse coefficients

At the decoder side, the dominant components, whose the energies are higher than the masking threshold, are obtained by adding the decoded residual component to the masking threshold. The remaining zero coefficients are then filled with the offset based on the spectral envelope.

First, the decoded residual coefficient $\hat{D}(k)$ is obtained from

$$\hat{D}(k) = \beta \hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor) \cdot \hat{D}^{\mathrm{norm}}(k), \quad (4)$$

where $\hat{D}^{\mathrm{norm}}(k)$ is the AVQ-decoded coefficient. $\beta \hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor)$ is the same value as in (3). Second, the decoded input MDCT coefficient $\hat{S}(k)$ is calculated as

$$\hat{S}(k) = \begin{cases} \mathrm{sgn}(\rho)\, \hat{f}'_{\mathrm{rms}}(\lfloor k/N \rfloor) & \text{if} \quad \hat{D}(k) = 0 \\ \mathrm{sgn}\left(\hat{D}(k)\right)\left\{\left|\hat{D}(k)\right| + \alpha \hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor)\right\} & \text{otherwise} \end{cases}, \quad (5)$$

where $\alpha \hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor)$ is the same value as in (2). ρ is a random value, and $\mathrm{sgn}(\rho)$ randomly outputs one or minus one. $\hat{f}'_{\mathrm{rms}}(\lfloor k/N \rfloor)$ is the offset based on the decoded spectral envelope; the zero coefficients are filled with the offset, and their signs will be randomly set to generate natural noise. The offset $\hat{f}'_{\mathrm{rms}}(\lfloor k/N \rfloor)$ is performed as follows:

$$\hat{f}'_{\mathrm{rms}}(\lfloor k/N \rfloor) = \sqrt{\frac{N\left[\hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor)\right]^2 - \sum_{i=k}^{k+N-1}\left[\hat{S}^{\mathrm{tmp}}(i)\right]^2}{\sum_{i=k}^{k+N-1} f_{\mathrm{zero}}(i)}}, \quad (6)$$

where i is the increment iteration number. The denominator represents the number of zero coefficients in $\hat{D}(k)$. $\hat{S}^{\mathrm{tmp}}(k)$ is the temporary reconstructed spectrum from $\hat{D}(k)$, represented as

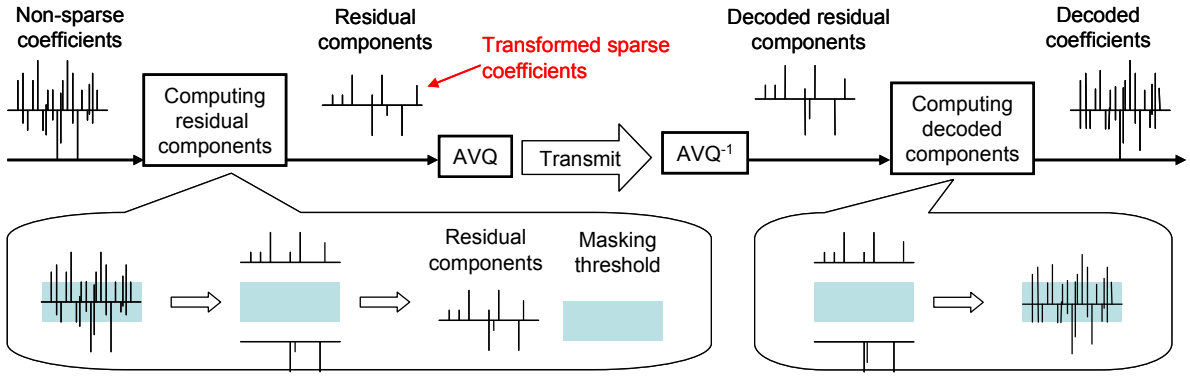

Figure 3: *Conceptual diagram of non-sparse mode.*

$$\hat{S}^{\mathrm{tmp}}(k)=\begin{cases}0 & \text{if} \quad \hat{D}(k)=0\\ \left|\hat{D}(k)\right|+\alpha\hat{f}_{\mathrm{rms}}(\lfloor k/N\rfloor) & \text{otherwise}\end{cases} \tag{7}$$

$f_{\mathrm{zero}}(i)$ is a flag to indicate the zero coefficients in $\hat{D}(k)$ as follows:

$$f_{\mathrm{zero}}(i)=\begin{cases}1 & \text{if } \hat{D}(i)=0\\ 0 & \text{otherwise}\end{cases} \tag{8}$$

3.2. Mode selector

The spectral masking can suppress the perceptual sound degradation in encoding the non-sparse signal using the sparse type of vector quantization. However the decoded output occasionally sounds like a noisy signal to listeners when the sparse signal is coded. In that case, the sparseness mode, which uses the ordinary AVQ, is more suitable for coding those sparse signals. Therefore, a method that switches the spectral masking on and off is also proposed for achieving high sound quality irrespective of the sound source.

The proposed mode selection is based on the frequency-domain sparseness analysis. Using the sparseness detection of input MDCT coefficients, the mode selector classifies the input signal into two coding modes: "sparse" mode and "non-sparse" mode. When the input MDCT coefficients are sparse, such as harmonic components of the music, the energy of those coefficients would be concentrated in the dominant components, Focusing on that fact, the coding-mode flag $b_{\mathrm{mode}}^{(m)}$ in the current frame m is computed as

$$b_{\mathrm{mode}}^{(m)}=\begin{cases}1, & \text{if } E[c]\le\delta_1\\ b_{\mathrm{mode}}^{(m-1)}, & \text{if } \delta_1<E[c]<\delta_2,\\ 0, & \text{otherwise}\end{cases} \tag{9}$$

$$c=\sum_{k=0}^{L-1}b_{\mathrm{sparse}}(k), \tag{10}$$

$$b_{\mathrm{sparse}}(k)=\begin{cases}1 & \text{if } \left|S^{\mathrm{norm}}(k)\right|<\varepsilon\\ 0 & \text{otherwise}\end{cases} \tag{11}$$

,

where $E[\cdot]$ is the average between the current and previous frames. δ_1 and δ_2 are the threshold values for deciding whether the spectrum is sparse. The current coding mode is decided depending on the previous flag when $\delta_1<E[c]<\delta_2$ in order to prevent frequent switching of the coding mode. c is the sparseness counter to numerate the number of frequency coefficients that have a smaller amplitude than the average value of the normalized coefficients. $b_{\mathrm{sparse}}(k)$ is a flag indicating whether the absolute value of $S^{\mathrm{norm}}(k)$ is lower than the threshold ε. One bit is transmitted as the information of the encoding-mode flag.

4. Implementation for G.711.1 D/G.722B

We implemented our method for SWB extension layers in ITU-T G.711.1 Annex D and ITU-T G.722 Annex B. An encoding block diagram of our method used in the codecs is shown in Figure 4. The codecs are based on an embedded scalable structure. The bit rates are extended to 96/112/128 kbit/s for G.711.1 Annex D and 64/80/96 kbit/s for G.722 Annex B. Both coders operate on 5-ms frames with the input and output sampled at 32 kHz.

The input signal is divided into two 16-kHz-sampled wideband and super higher-band (SHB, 8–16 kHz band) signals using a quadrature mirror filterbank (QMF). The wideband signal is divided into two 8-kHz-sampled lower band (LB, 0–4 kHz) and higher band (HB, 4–8 kHz) signals with the QMF. The LB and HB signals are coded with a core

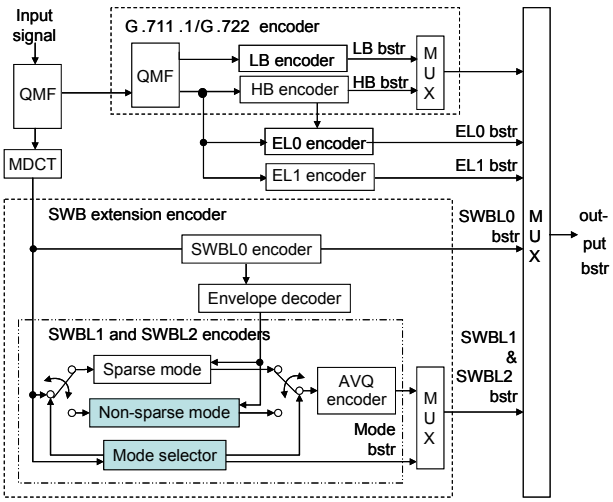

Figure 4: *Encoding block diagram of new method in G.711.1 Annex D and G.722 Annex B (bstr: bitstream).*

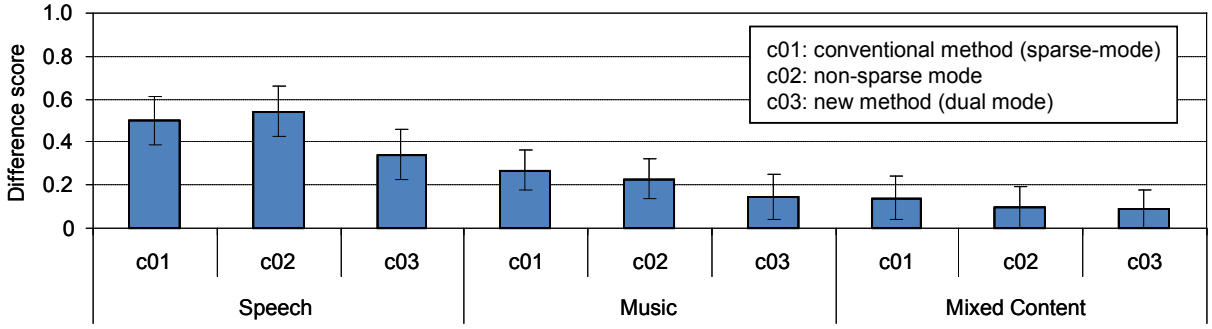

Figure 5: *Quality-difference score for speech, music, and mixed content.*

encoder (i.e., either G.711.1 or G.722) or EL0 and EL1. The SHB signal is coded in SWB extension sub-layers (SWBL0, SWBL1, and SWBL2).

4.1. SWB extension sub-layers

The SHB signal is coded in the MDCT frequency domain. The SWB expansion sub-layers consist of three sub-layers of SWBL0, SWBL1, and SWBL2. As more sub-layers are used, the audio quality improves. These sub-layers operate with 80 SHB MDCT coefficients. The first 64 coefficients, which are associated with the frequency range 8–14.4 kHz, are coded. The 64 coefficients are divided into 8 sub-bands, each with 8 coefficients.

In the SWBL0 sub-layer, the spectral envelope of the SHB MDCT coefficients is computed as a set of RMS values per sub-band and is vector-quantized. In the SWBL1 sub-layer, the three sub-bands that are perceptually important out of eight sub-bands are coded, and the remaining sub-bands are processed in the SWBL2 sub-layer. The information of the encoding mode is transmitted using one bit of the SWBL1 bitstream. In the SWBL1 and SWBL2 sub-layers, the encoding mode is selected according to the sparseness of the SHB MDCT coefficients. If the frame was classified as "sparse", the AVQ of an 8-dimensional vector is used to encode the SHB MDCT coefficients normalized per sub-band. Otherwise, the non-sparse mode is used to encode the residual coefficients using the AVQ.

5. Evaluation

The performance of the presented method as a part of G.711.1 Annex D was evaluated using a subjective listening test. The triple stimulus/hidden reference/double blind method ('Ref', 'A', 'B') with a five-point impairment scale, compliant with ITU-R BS.1116-1 [7], was used in the testing. Listeners evaluated three conditions: the output signals of the sparse mode, non-sparse mode, and new method. The above three conditions were implemented with the G.711.1 Annex D, μ-law, at 112 kbit/s and were evaluated for three sound items (speech, music, and mixed content) to confirm the effectiveness of the new method.

The test results comparing the conventional and new methods are shown in Figure 5 in the form of the difference scores, which represent the degradation of the coded output signal from the original signal. The vertical lines in the figure denote the 95% confidence interval. "c01: conventional method (sparse-mode)" is the ordinary AVQ, "c02: non-sparse mode" is the AVQ with proposed adaptive spectral masking, and "c03: new method (dual mode)" is the combinatorial method for the sparse mode and non-sparse mode, respectively. For each item, mean scores were given for 5 sound signals by 22 expert listeners. As these results show, the better scores were observed for all sound items by using the new method combing the sparse mode with the non-sparse mode. In particular, for speech, the new method improved the sound quality by about 0.2 points on a 5-point scale compared with the conventional method. On average, for the speech, music, and mixed content, the new method improved the sound quality by more than 0.1 points, and the significant improvement was confirmed..

6. Conclusions

A new coding method added to the Recommendations ITU-T G.711.1 Annex D and G.722 Annex B was described. An adaptive spectral masking of AVQ using a spectral envelope was proposed for MDCT-domain non-sparse signals. A method that switches the spectral masking on and off was proposed for achieving high sound quality irrespective of the sound source. Subjective evaluations showed that the sound qualities were improved by about 0.2 points (on a 5-point scale) for speech and by more than 0.1 points on average for speech, music and mixed content, and the significance of the improvement was validated.

7. References

[1] M. Xie and J.-P. Adoul, "Embedded algebraic vector quantization (EAVQ) with application to wideband audio coding," In Proc. IEEE ICASSP, Atlanta, GA, vol. 1, pp. 240–243, May 1996.

[2] R. M. Gray, "Vector quantization," IEEE ASSP Mag., vol. I, pp. 4–29, April 1984.

[3] ITU-T (Telecommunication Standardization Sector, International Telecommunication Union), Geneva, Switzerland, ITU-T G.711.1 Annex D (pre-published), – New Annex D with superwideband extension, Nov. 2010.

[4] ITU-T (Telecommunication Standardization Sector, International Telecommunication Union), Geneva, Switzerland, ITU-T G.722 Annex B (pre-published), – New Annex B with superwideband extension, Nov. 2010.

[5] ITU-T (Telecommunication Standardization Sector, International Telecommunication Union), Geneva, Switzerland, ITU-T G.711.1 – Wideband embedded extension for G.711 pulse code modulation, Mar. 2008.

[6] ITU-T (Telecommunication Standardization Sector, International Telecommunication Union), Geneva, Switzerland, ITU-T G.722 – 7 kHz audio-coding within 64 kbps, Nov. 1988.

[7] ITU-R BS.1116-1, "Method for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1997.