IEICE TRANS. INF. & SYST., VOL.E97–D, NO.5 MAY 2014

PAPER

Adaptive Spectral Masking of AVQ Coding and Sparseness Detection for ITU-T G.711.1 Annex D and G.722 Annex B Standards

Masahiro FUKUI[†a)], **Shigeaki SASAKI**[††], **Yusuke HIWASAKI**[†], ***Members***, **Kimitaka TSUTSUMI**[†††], **Sachiko KURIHARA**[†], ***Nonmembers***, **Hitoshi OHMURO**[†], ***Member***, ***and*** **Yoichi HANEDA**[††††], ***Nonmember***

SUMMARY We proposes a new adaptive spectral masking method of algebraic vector quantization (AVQ) for non-sparse signals in the modified discreet cosine transform (MDCT) domain. This paper also proposes switching the adaptive spectral masking on and off depending on whether or not the target signal is non-sparse. The switching decision is based on the results of MDCT-domain sparseness analysis. When the target signal is categorized as non-sparse, the masking level of the target MDCT coefficients is adaptively controlled using spectral envelope information. The performance of the proposed method, as a part of ITU-T G.711.1 Annex D, is evaluated in comparison with conventional AVQ. Subjective listening test results showed that the proposed method improves sound quality by more than 0.1 points on a five-point scale on average for speech, music, and mixed content, which indicates significant improvement.

key words: *speech and audio coding, standardization, ITU-T G.711.1 Annex D, ITU-T G.722 Annex B, super-wideband (SWB) extension, algebraic vector quantization (AVQ)*

1. Introduction

An increasing number of telecommunication services, e.g., for video conferencing and videophones, are being used over broadband networks. For better sound clarity and less listener fatigue, some services offer frequency ranges approaching near-CD-audio quality, such as the 14-kHz super-wideband (SWB).

Speech coding is one of the key technologies in telecommunication systems. It is necessary for speech transmission even over broadband networks. 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. To meet the demand for these SWB codecs, several standardization activities have been conducted recently.

Manuscript received August 22, 2013.
Manuscript revised December 20, 2013.

[†]The authors are with NTT Media Intelligence Laboratories, NTT Corporation, Musashino-shi, 180–8585 Japan.
[††]The author is with NTT Advanced Technology Corporation, Kawasaki-shi, 210–0007 Japan.
[†††]The author is with Research Laboratories, NTT DOCOMO, INC., Yokosuka-shi, 239–8536 Japan.
[††††]The author is with the Faculty of Informatics and Engineering, The University of Electro-Communications, Chofu-shi, 182–8585 Japan.
a) E-mail: fukuimas@ieee.org
DOI: 10.1587/transinf.E97.D.1264

A fast and efficient coding method is important for processing a larger number of 32-kHz sampled signals. One well-known coding method is algebraic vector quantization (AVQ) [1]–[3], which is a fast searching method of vector quantization [4], [5]. When the linear predictive coding (LPC) residual of a speech signal is sparse, AVQ is likely to select the sparse codevectors which are represented with only a few nonzero components. The sparse codevectors can also be applied to the frequency domain coding with modified discrete cosine transform (MDCT) [6] because sparseness can be found in some music sound items such as string instruments, as well as in speech signals in that domain. For quantizing "non-sparse" frequency-domain signals (e.g., noise components included in music), AVQ can select the full pulse codevectors under many bit budgets. However, the lack of available bit budget occasionally causes the perceptual degradation of sound quality when quantizing those kinds of signals.

To solve this problem, an adaptive spectral masking of AVQ for non-sparse signals in the MDCT domain 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 introduced to ITU-T Recommendations G.711.1 Annex D [7], [8] and G.722 Annex B [8], [9] (SWB extensions of G.711.1 [10]–[12] and G.722 [13], [14]), which were standardized in November 2010.

ITU-T launched the work item in October 2007, and the Terms of References (ToRs) for these two codec standards were finalized and approved in April 2008. A qualification phase was first conducted to check whether candidates could pass all requirements. This was followed by the optimization and characterization phase, where five organizations (ETRI, France Telecom, Huawei Technologies, VoiceAge, and NTT) constructively collaborated on creating a unified algorithm. NTT proposed and introduced their new technologies described in this paper, and integrated the technologies from other four organizations into new codecs. NTT also contributed to the Consortium by helping to run its operations and drive its activities. This paper presents overviews of the codecs, which are the outcomes of the collaboration effort.

The remainder of this paper is organized as follows.

Section 2 describes the conventional coding scheme with AVQ. Section 3 presents AVQ with new spectral masking and the switching method of the spectral masking with sparseness detection. Section 4 provides the details of the implementation of our method in ITU-T G.711.1 Annex D and ITU-T G.722 Annex B. Section 5 describes subjective evaluation results, and Sect. 6 concludes the paper.

2. Coding Scheme with AVQ

An example encoding block diagram of conventional AVQ in the frequency domain is shown in Fig. 1. The input signal is transformed into the MDCT coefficient $S(k)$, where $k = 0, \cdots, L-1$, and L is the number of samples in the frame; L was set as 80. The L MDCT coefficients are split into each N sample of M sub-bands; in this case, $N = 8$ and $M = 10$. 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 $S(k)$ is normalized using the quantized spectral envelope $\hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor)$ as

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

and the obtained normalized coefficient $S'(k)$ is encoded using AVQ.

3. New Spectral Masking and Sparseness Detection

As described in Sect. 1, when the available bit budget is insufficient to quantize non-sparse signals such as the noise components in music, the conventional coding scheme occasionally causes sound quality degradation. This problem is addressed in the proposed method by switching and using efficient coding methods for both the sparse and non-sparse signals based on the results of the sparseness analysis of the input MDCT coefficient.

The encoding block diagram of the proposed method is shown in Fig. 2. The conventional scheme with AVQ, described in Sect. 2, is used in "sparse mode", and the new non-sparse type coding is used in "non-sparse mode". Given the input MDCT coefficient, $S(k)$, the mode selector determines the encoding mode according to the sparseness of $S(k)$. If the sparse mode is selected, AVQ encodes the normalized coefficients in the MDCT domain, as conventionally done. Otherwise, in the non-sparse mode, only the coefficients with energies higher than the masking threshold are encoded. The non-sparse mode and the mode selector are presented in detail in the following subsections.

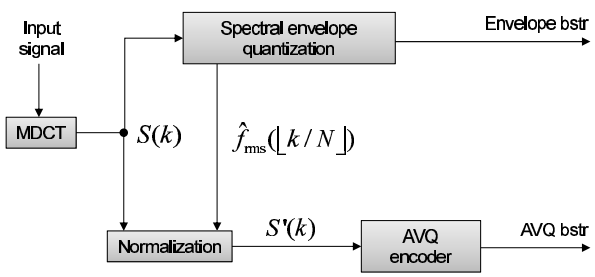

Fig. 1 Encoding block diagram of conventional AVQ (bstr: bitstream).

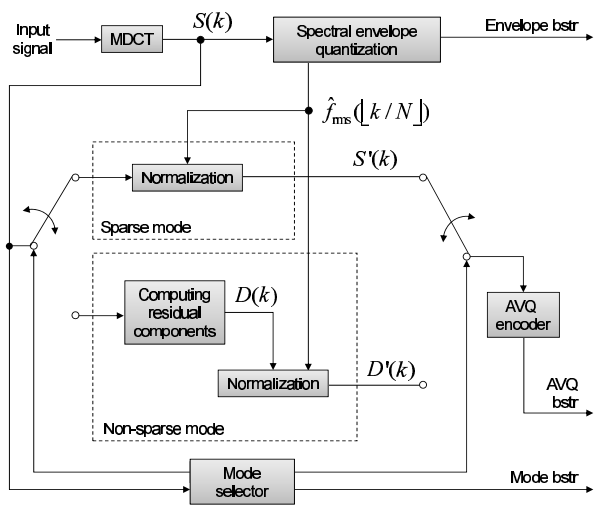

Fig. 2 Encoding block diagram of proposed method (bstr: bitstream).

3.1 Non-sparse Mode

With non-sparse signals, zero sequences in the decoded coefficients can be easily detected as audible degradations. The non-sparse mode aims at reducing this perceptual sound degradation by filling the non-dominant components below the masking threshold with random noise. A conceptual diagram of the non-sparse mode is shown in Fig. 3. On the encoder side, when the input coefficients are classified as non-sparse, they are transformed into the sparse residual components in the following procedure; the method subtracts the masking threshold from the absolute values of non-sparse coefficients and truncates the components below the masking threshold towards zero. The masking threshold is adaptively determined per the sub-band in order to prevent the residual components from becoming dense. The proposed method maintains the degree of sparseness of the residual components by using the spectral envelope of the non-sparse coefficients, which is the RMS value per the sub-band, as a criterion of the masking threshold. The obtained residual components are vector-quantized using AVQ. The decoder side reconstructs the non-sparse coefficients by adding the decoded residual components to the offset based on the masking threshold.

3.1.1 Computing and Encoding Residual Components

The basic concept for computing residual components is illustrated in Fig. 4. The constant parameters used in the non-sparse mode are listed in Table 1. The frequency components, of which the amplitudes of the inputs, $S(k)$, are higher than the masking threshold, are converted into the differences between their magnitude and masking threshold, and the rest of the components are set to zero. As a result, the obtained residual components are composed mainly of zero

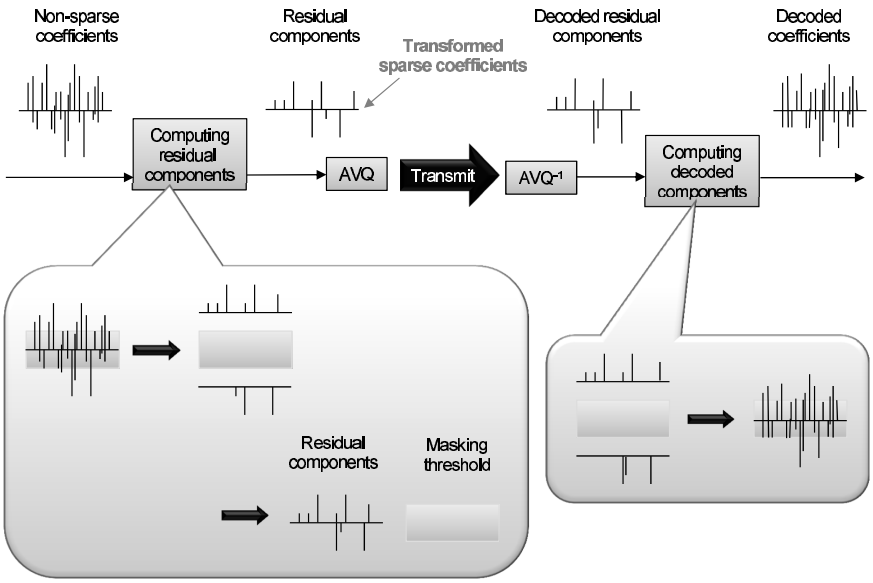

Fig. 3 Conceptual diagram of non-sparse mode.

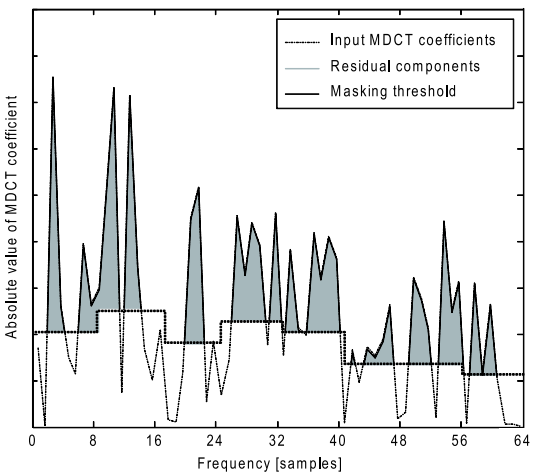

Fig. 4 Basic concept for computing residual components.

Table 1 Constant parameters used in non-sparse mode.

Parameter	Value
α	0.5
β	0.6

and a few non-zero components. These non-zero components are dominant in the input coefficients. Since the residual components are sparse compared to the inputs, $S(k)$, they are effectively encoded using the sparse type of vector quantization, such as AVQ.

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

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

where sgn[·] is the polarity of $S(k)$, max(,) is the maximum value selection, $\alpha \hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor)$ is the masking threshold used for calculating the sparse residual components, and α is the factor for adjusting the sparseness of the residual components. The factor α is set as the constant here in order to avoid the quantization and transmission of α. The larger the factor α is, the larger the noise-masking effect is. However, if α is too large, the non-sparse mode degrades the quality of the decoded signal even when the signal is non-sparse. Therefore, it is important to choose the optimal α. The value of α was determined by verifying the performance for 12 sound items such as pop music, rock music, classic music, and a news jingle.

The obtained error spectrum from Eq. (2) is normalized as follows:

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

The RMS of $D(k)$ for normalization is not transmitted, so $\beta \hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor)$ is used in place of the RMS of $D(k)$ in Eq. (3). The term β is the constant factor to approximate $\hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor)$ as the RMS value of $D(k)$, which is less than or equal to one. The factor β was determined in a preliminary experiment, as in the case with factor α. The normalized coefficient $D'(k)$ is encoded using AVQ and transmitted to the decoder.

3.1.2 Reconstructing Non-sparse Coefficients

At the decoder side, the dominant components, whose 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 component $\hat{D}(k)$ is obtained from

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

where $\hat{D}'(k)$ is the AVQ-decoded coefficient, and

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

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

where $\mathrm{sgn}\left[\hat{D}(k)\right]$ is the polarity of $\hat{D}(k)$, $\alpha\hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor)$ is the same value as in Eq. (2), ρ is a random value, $\mathrm{sgn}[\rho]$ is a random polarity with an output of one or minus one, and $\hat{f}'_{\mathrm{rms}}(\lfloor k/N \rfloor)$ is the offset based on the decoded spectral envelope. If $\hat{D}(k)$ has a zero coefficient, the coefficient is filled with the offset, and the sign will be randomly set in order to generate natural noise. The calculation of offset $\hat{f}'_{\mathrm{rms}}(\lfloor k/N \rfloor)$ is performed as follows:

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

where i is the increment iteration number and $f_z(i)$ is a flag to indicate the zero coefficients in $\hat{D}(k)$ as follows:

$$f_z(i) = \begin{cases} 1 & \text{if } \hat{D}(i) = 0 \\ 0 & \text{otherwise} \end{cases}. \tag{7}$$

The denominator of Eq. (6) represents the number of zero coefficients in $\hat{D}(k)$, and $\hat{S}_\tau(i)$ is the temporary reconstructed spectrum from $\hat{D}(k)$, represented as

$$\hat{S}_\tau(i) = \begin{cases} 0 & \text{if } \hat{D}(k) = 0 \\ |\hat{D}(k)| + \alpha\hat{f}_{\mathrm{rms}}(\lfloor k/N \rfloor) & \text{otherwise} \end{cases}. \tag{8}$$

The numerator of Eq. (6) is approximately equal to the amplitude of the input $S(s)$ by computing the difference between total powers at input $S(s)$s and at the temporary reconstructed spectrum from $\hat{D}(k)$. The offset $\hat{f}'_{\mathrm{rms}}(\lfloor k/N \rfloor)$ represents a set of RMS values of the inputs, $S(s)$, when $\hat{D}(k)$ is a zero coefficient.

3.2 Mode Selector

The spectral masking can suppress 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 this case, the sparse mode, which uses the conventional coding scheme with AVQ, is more suitable for coding those sparse signals. Therefore, a method for switching the spectral masking on and off is also proposed for achieving high sound quality irrespective of the sound source.

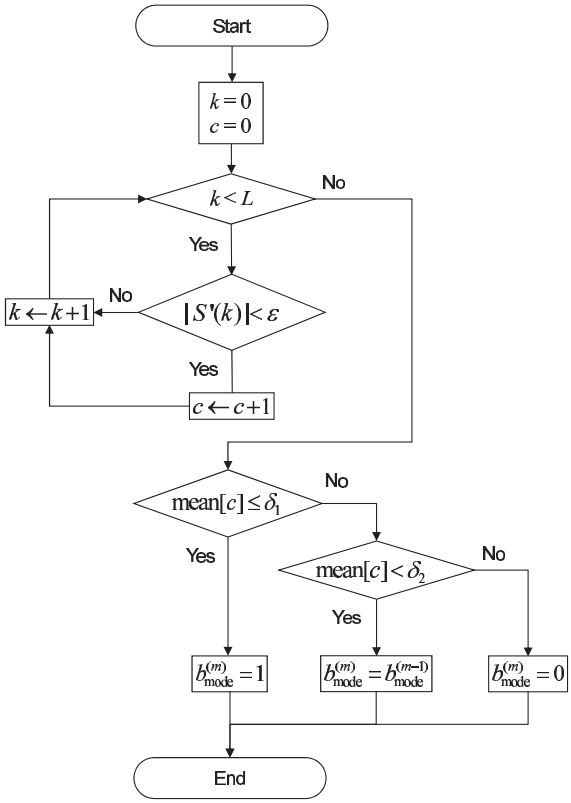

Fig. 5 Flowchart of mode selector.

Table 2 Constant parameters used in mode selector.

Parameter	Value
ϵ	0.5
δ_1	15
δ_2	20

A flowchart of the mode selector is shown in Fig. 5. The constant parameters used in the mode selector are listed in Table 2. Based on the preliminary experiments, the parameters were selected in order to achieve high sound quality irrespective of the sound source for 12 sound items such as pop music, rock music, classic music, and a news jingle. Our mode selector is based on frequency-domain sparseness analysis. The mode selector uses the sparseness detection of inputs, $S(k)$, per frame to classify the input signal into two coding modes: sparse and non-sparse. When the inputs, $S(k)$, are sparse, such as with harmonic components of the music, the energy of those coefficients are concentrated in the dominant components, as shown in Fig. 6. Focusing on this fact, we compute the coding-mode flag $b^{(m)}_{\mathrm{mode}}$ in the current frame m as

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

where mean[·] is the average between the current and previous frames. If the inputs, $S(k)$, are categorized as non-

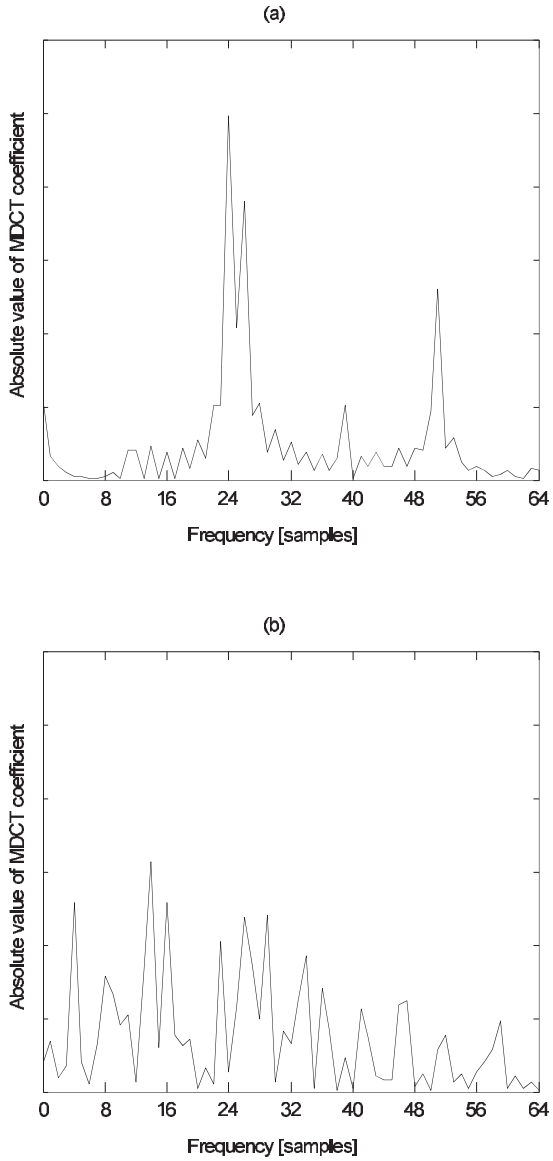

Fig. 6 (a) Spectrum of sparse signal (single strings) and (b) spectrum of non-sparse signal (rock music).

sparse, then $b_{\text{mode}}^{(m)} = 1$; otherwise, $b_{\text{mode}}^{(m)} = 0$. The initial value of the coding-mode flag is set to zero. One bit is used as the information of the coding-mode flag and is transmitted.

The terms δ_1 and δ_2 in Eq. (9) are the threshold values for determining whether the spectrum is sparse. The current coding mode is determined depending on the previous flag when $\delta_1 < \text{mean}[c] < \delta_2$ in order to prevent frequent switching of the coding mode. The term 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 as follows:

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

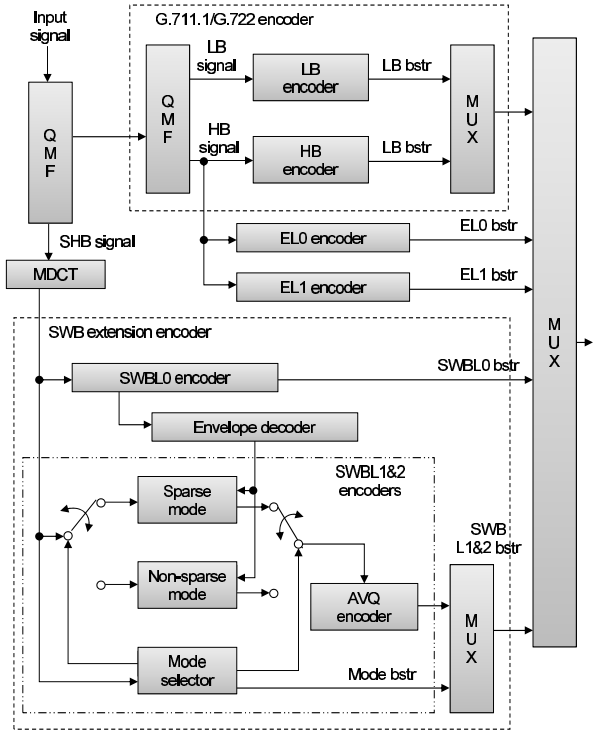

Fig. 7 Encoding block diagram of proposed method in G.711.1 Annex D and G.722 Annex B (bstr: bitstream).

$$b_{\text{sparse}}(k) = \begin{cases} 1 & \text{if } |S'(k)| < \epsilon \\ 0 & \text{otherwise} \end{cases}, \tag{11}$$

where $b_{\text{sparse}}(k)$ is a flag indicating whether the absolute value of $S'(k)$ is lower than the threshold ϵ.

4. Overviews of New Codecs and Implementation of Proposed Method

This section presents two summaries: the codec overview of new standards ITU-T G.711.1 Annex D and G.722 Annex B and the implementation of the proposed method for these new codecs.

4.1 Codec Overview

An encoding block diagram of Recommendations ITU-T G.711.1 Annex D and G.722 Annex B is shown in Fig. 7. These codecs are based on an embedded scalable structure and 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 encoder (i.e., either G.711.1 or G.722) or HB enhancement sub-layers of EL0 and EL1. Finally, the SHB signal is coded in the SWB extension sub-layers of SWBL0, SWBL1, and SWBL2.

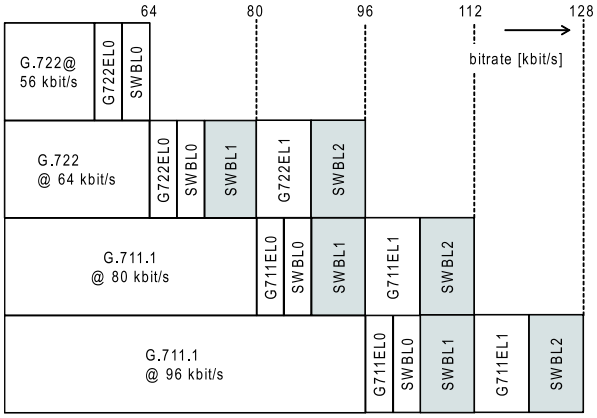

Fig. 8 Layered structure of G.711.1 Annex D and G.722 Annex B bitstreams.

The layered structure of G.711.1 Annex D and G.722 Annex B bitstreams is shown in Fig. 8. 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.

4.1.1 HB Enhancement Sub-Layers

The HB enhancement sub-layers, which are the EL0 and EL1 encoders shown in Fig. 7, are the two codec dependent sub-layers (G711EL0/EL1 and G722EL0/EL1) for further enhancing the HB signal coding. The details of these sub-layers are as follows:

- G711EL0/EL1
  In the first enhancement sub-layer, G711EL0, the 6.4-8.0-kHz part of the HB signal is coded using AVQ in the MDCT frequency domain because the highest frequency covered by the G.711.1 core codec is 7.0 kHz. In the subsequent G711EL1 sub-layer, the residual coefficients, i.e., the difference between the HB signal and the decoded signal from the G.711.1 core and G711EL0, are coded using dynamic bit allocation and scalar quantization.
- G722EL0/EL1
  The first G722EL0 sub-layer enhances the 19 most perceptually important HB time samples (out of 40) using scalable scalar quantization. The subsequent G722EL1 sub-layer refines the whole HB frame of 40 samples using an additional scalar quantization. In addition to the scalable quantization of G722EL0 and G722EL1, the LB part of the legacy G.722 encoder is enhanced using a noise feedback loop to perceptually shape the ADPCM coding noise.

4.1.2 SWB Extension Sub-Layers

The SHB signal is coded in the MDCT frequency domain. The SWB expansion sub-layers consist of the 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 then vector-quantized. In the SWBL1 sub-layer, the three sub-bands that are perceptually important out of the eight sub-bands are coded, and the remaining sub-bands are processed in the SWBL2 sub-layer.

Table 3 Worst-case complexities.

Processor name	Complexity [WMOPS]
Sparse mode	2.76
Non-sparse mode	3.54
Mode selector	0.05

4.2 Implementation for ITU-T G.711.1D and G.722B

Our method was added for SWB extension layers in ITU-T G.711.1 Annex D and ITU-T G.722 Annex B. In the SWBL1 and SWBL2 sub-layers, the encoding mode is selected according to the sparseness of the SHB MDCT coefficients; the encoding mode information, which is the coding-mode flag described in Sect. 3.2, is transmitted using one bit of the SWBL1 bitstream. If the frame is classified as "sparse", 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 components, which are calculated as described in Sect. 3.1.1, using the same AVQ as that of the sparse mode.

Table 3 lists the processing complexities of sparse mode, non-sparse mode, and mode selector. The complexities are expressed as the worst-case complexity in weighted millions operations per second (WMOPS); the worst-case complexity is the performance figure for real-time system. These are based on complexity reports using the basic operators, which simulate DSP opearations, of ITU-T Software Tool Library (STL2005) in ITU-T G.191 [15]. They are implanted into C-source codes, the codes including those are compiled and then the WMOPS is obtained by executing the compiled binary file. As indecated in Table 3, the complexity of the proposed method (non-sparse mode and mode selector) remains at the slightly increased level of about 0.8 WMOPS compared with that of the conventional method (sparse mode).

5. Evaluation

5.1 Experimental Conditions

Most phase characteristics of the decoded signal are discordant from those of the input signal in the proposed coding method, so it is difficult to obtain objective speech quality that has a correlation with subjective speech quality using an objective evaluation method such as signal-to-noise ratio (SNR). Therefore, the performance of the proposed method

Table 4 Experimental conditions.

Methodology	ITU-R BS.1116-1
Sampling rate	32 kHz
Frame length	5 ms
Frame shift	5 ms
MDCT points	80 samples
# of listeners	22

Table 5 Overview of test signals.

Sound item	Languages	Type of signals
Clean speech	Japanese	Female talker
		Female talker
		Female talker
		Male talker
		Male talker
Music	Japanese	Pop music
		Rock music
		Vocal music
		Classical music
		Jazz music
Mixed content	Japanese	TV commercial message
		TV commercial message
		TV program
		News artificially mixed with jingles
		News artificially mixed with jingles

Table 6 List of conditions.

Number	Test condition	Level [dBov]
c00	Direct [50-14000 Hz]	−26
c01	Conventional method (sparse mode)	−26
c02	Non-sparse mode	−26
c03	Proposed method (dual mode)	−26

as a part of G.711.1 Annex D was evaluated using a subjective listening test. The experimental conditions, overview of signals, and list of conditions are listed in Tables 4, 5, and 6, respectively. The triple stimulus/hidden reference/double blind method ('Ref', 'A', 'B') with a five-point impairment scale, compliant with ITU-R BS.1116-1 [16], was used in the testing. ITU-R BS.1116-1 is aimed at determining whether a test signal contains an audible distortion compared with a reference. The results are represented as quality-difference scores, which are the differences in mean opinion score (MOS) between the reference and evaluation signals. In BS.1116-1, the MOS of the decoded signal is expected to be graded on a scale of 1.0 to 4.9 points, and that of the original signal is set to 5.0 points. Therefore, the difference score in BS.1116-1 is equal to the absolute quality. All reference and evaluation signals were presented to both ears via headphones (Sennheiser HD 280 Pro). Listeners evaluated three conditions: the output signals of the sparse mode, non-sparse mode, and proposed method. The above three conditions were implemented with the G.711.1 Annex D using μ-law core at 112 kbit/s and were evaluated for three sound items (speech, music, and mixed content) to evaluate the effectiveness of the proposed method. The mixed content consisted of speech plus music and/or background noise.

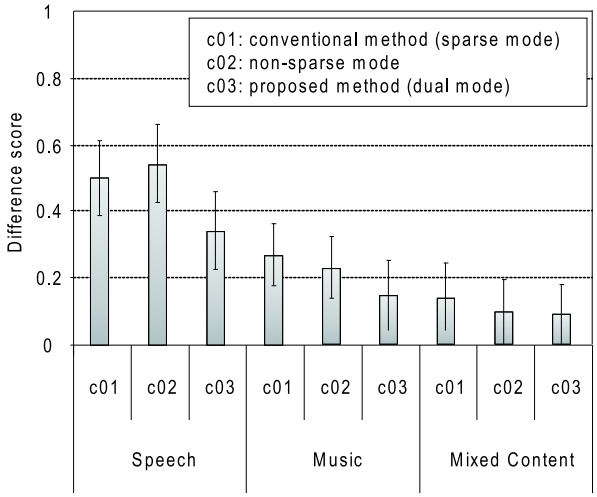

Fig. 9 Quality-difference scores between the original and coded output signals for speech, music, and mixed content.

5.2 Experimental Results

The test results comparing the conventional coding scheme with AVQ and the proposed method are shown in Fig. 9 in the form of 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, and "c01: conventional method (sparse-mode)" is the conventional scheme with AVQ, "c02: non-sparse mode" is AVQ with the proposed adaptive spectral masking, and "c03: proposed method (dual mode)" is the combinatorial method for the sparse and non-sparse modes. For each item, 22 experienced listeners (8 males and 14 females ranging in age from 21 to 36) gave mean scores for five sound signals.

As these results show, better scores were observed for all sound items by using the proposed method that combines the sparse and non-sparse modes. In particular, the proposed method improved the sound quality for speech by about 0.2 points on a five-point scale compared with the conventional method. The significant improvement for speech can be observed from point of view of that the confidence interval of the proposed method is below the average of the conventional method. On the other hand, the score of the non-sparse mode for speech was the lowest for all modes. This is because the quantization error of the non-sparse mode was high when coding sparse signals such as vowel sounds. For the mixed content, the difference in scores between the non-sparse mode and the proposed method was small; but this is because almost all periods of input signals were categorized as non-sparse. As can be seen from these results, the proposed method improved the sound quality by more than 0.1 points on average for all sound items, and achieved high sound quality irrespective of the sound source.

6. Conclusion

A coding method introduced into Recommendations ITU-T

G.711.1 Annex D and G.722 Annex B was proposed. An adaptive spectral masking of AVQ using a spectral envelope was proposed for non-sparse signals in the MDCT domain. This method switches the spectral masking on and off as dual-mode coding in order to achieve high sound quality irrespective of the sound source. Subjective experiments revealed that the proposed coding method outperformed the conventional coding scheme with AVQ, and it improved sound quality by more than 0.1 points on a five-point scale on average for speech, music, and mixed content. In particular, the proposed method produced significant improvement for speech signals, which improved the sound quality of speech by about 0.2 points. This result shows that the proposed dual-mode coding appropriately quantized both the sparse components such as vowel and the non-sparse components.

References

[1] ITU-T (Telecommunication Standardization Sector, International Telecommunication Union), Geneva, Switzerland, ITU-T G.718 – Frame error robust narrowband and wideband embedded variable bit-rate coding of speech and audio from 8-32 kbit/s, June 2008.
[2] T. Vaillancourt et al., "ITU-T EV-VBR: A robust 8-32 kbit/s scalable coder for error prone telecommunications channels," EUSIPCO, Aug. 2008.
[3] S. Ragot, B. Bessette, and R. Lefebvre, "Low-complexity multi-rate lattice vector quantization with application to wideband TCX speech coding at 32 kbit/s," Proc. ICASSP2004, vol.1, pp.501–504, May 2004.
[4] R.M. Gray, "Vector quantization," IEEE ASSP Mag., vol.I, pp.4–29, April 1984.
[5] G. Allen and R.M. Gray, Vector Quantization and Signal Compression, Kluwer Academic Publishing, 1991.
[6] J.P. Princen and A.B. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-34, no.5, pp.1153–1161, Oct. 1986.
[7] ITU-T (Telecommunication Standardization Sector, International Telecommunication Union), Geneva, Switzerland, ITU-T G.711.1 Annex D – New Annex D with superwideband extension, Nov. 2010.
[8] L. Miao et al., "G.711.1 Annex D and G.722 Annex B - New ITU-T superwideband codecs," Proc. ICASSP2011, vol.7, pp.5232–5235, May 2011.
[9] ITU-T (Telecommunication Standardization Sector, International Telecommunication Union), Geneva, Switzerland, ITU-T G.722 Annex B – New Annex B with superwideband extension, Nov. 2010.
[10] ITU-T (Telecommunication Standardization Sector, International Telecommunication Union), Geneva, Switzerland, ITU-T G.711.1 – Wideband embedded extension for G.711 pulse code modulation, March 2008.
[11] Y. Hiwasaki and H. Ohmuro, "ITU-T G.711.1: Extending G.711 to higher-quality wideband speech," IEEE Commun. Mag., vol.47, no.10, pp.110–116, Oct. 2009.
[12] Y. Hiwasaki et al., "G.711.1: A wideband extension to ITU-T G.711," Proc. EUSIPCO2008, Lausanne, Switzerland, Aug. 2008.
[13] ITU-T (Telecommunication Standardization Sector, International Telecommunication Union), Geneva, Switzerland, ITU-T G.722 – 7 kHz audio-coding within 64 kbps, Nov. 1988.
[14] X. Maitre, "7 kHz audio coding within 64 kbit/s," IEEE Sel. Areas. Commun., vol.6, no.2, pp.283–298, Feb. 1988.
[15] ITU-T (Telecommunication Standardization Sector, International Telecommunication Union), Geneva, Switzerland, ITU-T G.191 – ITU-T Software Tool Library 2005 User's manual, Aug. 2005.
[16] ITU-R Rec. BS.1116-1, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1997.

Masahiro Fukui received the B.E. degrees in information science from Ritsumeikan University, Shiga, Japan, in 2002. He received the M.E. degree in information science from Nara Institute of Science and Technology, Nara, Japan, in 2004. Since joining Nippon Telegraph and Telephone Corporation (NTT) in 2004, he has been engaged in research field of acoustic echo canceller and speech coding. He is now a Research Engineer at NTT Media Intelligence Laboratories. He is a member of IEEE, IEICE and ASJ.

Shigeaki Sasaki is a Senior Manager, NTT Advanced Technology Corporation. He received the B.E. degree in physics from Kyoto University, Kyoto, in 1991. He joined NTT in 1991 and has been engaged in research on wideband speech coding. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan and the Acoustical Society of Japan (ASJ). He received the Achievement Award from IEICE in 2009 and the Maejima Hisoka Award from Tsushinbunka Association in 2010.

Yusuke Hiwasaki obtained B.E. in instrumentation engineering, M.E., and Ph.D. degrees in computer science, from Keio University, Yokohama, Japan, in 1993, 1995, and 2006, respectively. Since joining NTT in 1995, he has been engaged in research field of low bit-rate speech coding and voice-over-IP telephony. From 2001 to 2002, he was a guest researcher at Kungliga Tekniska Hogskolan (Royal Institute of Technology) in Sweden. He now works as a senior research engineer, supervisor, at NTT. Since 2007, he has been active in standardization of speech coding especially at ITU-T Study Group 16 (SG16) and acted as the editor of Recommendation ITU-T G.711.1. From 2009, he had been Associate Rapporteur of ITU-T SG16 Q.10, a question on speech coding matters, and then became Rapporteur in 2011. Dr. Hiwasaki received Technology Development Award from the Acoustical Society of Japan, Best Paper Award from IEICE Communications Society, IEICE Achievement Award, Teishin Association Maejima Award in 2006, 2006, 2009 and 2010, respectively. He is a member of IEEE, IEICE, and Acoustical Society of Japan.

Kimitaka Tsutsumi received the B.E., and M.E. degrees from Kyoto University, Kyoto, Japan in 2004, and 2006. From 2006 to 2010, he was with the Nippon Telegraph and Telephone Corporation (NTT), Japan. In 2010, he joined NTT DOCOMO, where he has been engaged in the research and development on speech coding, and standardization activities in 3GPP. He is a member of the Acoustical Society of Japan (ASJ).

Sachiko Kurihara is a Research Engineer, NTT Media Intelligence Laboratories. She joined Nippon Telegraph and Telephone Corporation (NTT) in 1985. In 1990, she graduated in electronics from the Junior College of the University of Electro-Communications. Since she joined NTT, she has been engaged in work on quality assessment of telephone calls. Later, she was engaged in research on speech coding and its quality and ITU speech coding standardization. She received Telecommunication Systems Technology prize awarded by the Telecommunications Advancement Foundation in 1996; the Acoustical Society of Japan Techonology Research and Development prize; and in 1998 the Director General Prize of Science and Technology Agency for Originality, Ingenuity, and Meritorious service in 2009, she received the Encouragement Prize by the Promotion Foundation for Electrical Science and Engineering. She is a member of the Acoustical Society of Japan (ASJ).

Hitoshi Ohmuro received the B.E. and M.E. degrees in electrical engineering from Nagoya University, Aichi, in 1988 and 1990, respectively. He joined NTT in 1990. He has been engaged in research on highly efficient speech coding and the development of VoIP applications. He is now a Senior Research Engineer, Supervisor at NTT Media Intelligence Laboratories. He is a member of IEEE, IEICE and ASJ.

Yoichi Haneda received the B.S., M.S., and Ph.D. degrees from Tohoku University, Sendai, in 1987, 1989, and 1999. From 1989 to 2012, he was with the Nippon Telegraph and Telephone Corporation (NTT), Japan. In 2012, he joined the University of Electro-Communications, where he is a Professor. His research interests include modeling of acoustic transfer functions, microphone arrays, loudspeaker arrays, and acoustic echo cancellers. He received paper awards from the ASJ and from the IEICE of Japan in 2002. Dr. Haneda is a senior member of the IEEE, and a member of ASJ.