# An Embedded Variable Bit-Rate Audio Coder for Ubiquitous Speech Communications

Do Young Kim[1] and Jong Won Park[2]

[1] Multimedia Communications Team, Electronics and Telecommunications Research Institute, 161 Gajeong-Dong, Yuseong-Gu, Daejeon, Rep. of Korea 305-700
dyk@etri.re.kr

[2] Department of Information Communications Engineering, Chungnam National University, 220 Gung-Dong, Yuseong-Gu, Daejeon, Rep. of Korea 305-764
jwpark@cnu.ac.kr

**Abstract.** In this paper, we propose an embedded variable bit-rate (VBR) audio coder to provide the fittest quality of service (QoS) and better connectivity of service for the ubiquitous speech communications. It has scalable bandwidth for narrowband to wideband speech signal, and embedded 8 32 kbit/s VBR corresponding to the network condition and terminal capacity. For the design of the embedded VBR coder, the narrowband signals are compressed by an existing standard speech coding method for the compatibility with G.729 coder, and then the other signals are compressed hierarchically on the basis of CELP enhancement and transform coding with temporal noise shaping (TNS) method. By the objective and subjective quality tests, it is shown that the proposed embedded VBR audio coder provides a reasonable quality compared with existing audio coders such as G.722 and G.722.2 in terms of mean opinion score (MOS) and perceptual evaluation of speech quality of wideband (PESQ-WB).

**Keywords:** Embedded Coder, G.729EV, MOS, PESQ-WB, Scalable Audio Coder, Ubiquitous Audio.

## 1 Introduction

The speech communications over the ubiquitous environment require better QoS than the existing telephony, better connectivity of service among various kinds of end points over the network, and the interoperability with the existing speech terminals. In this paper, we propose an embedded VBR audio coder so as to cope with the above requirements mainly for the ubiquitous speech communications. To meet the requirements, we consider the wideband speech coder which covers the full energy of human speech, and the embedded VBR audio coder in order to adapt its bit-rates dynamically from 8 to 32 kbit/s according to the variation of network especially between the different wireless networks and the capacity of the remote terminals [1]. Moreover, for backward compatibility with the popular narrowband speech coder used in the existing network, the proposed audio coder has a standard G.729 speech coder [2], [3], [4].

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 30–39, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Following this introduction, we will review the embedded VBR coder model that has been developed in the current state-of-the-art digital communication networks in Section 2. In Section 3, we will describe the structure of the proposed audio coder. In Section 4, we will evaluate the performance of the proposed audio coder by using the objective and subjective quality test method. Finally, we will present our conclusions in Section 5.

## 2 Embedded Audio Coder Model

Audible frequency range of human voice is from 20 to 20000 Hz. This audible frequency range can be divided into three parts, and we define these bands as narrowband (300∼3400Hz), wideband (50∼7000Hz), and audio band (20∼20000Hz). Because the energy of human speech is generally located in narrowband, narrowband speech coders have been developed and used. These speech coders started in PCM method [5], currently 8 kbps CS-ACELP which is standardized as ITU-T Recommendation G.729 is widely used [2], [3], [4]. With the advancement of network technology and internet service, many users have demanded the higher quality services and the research for wideband speech coder has been advanced. At present, G.722 [6] and G.722.2 [7] are widely used. These wideband coders have good performance for speech signals, but these cannot provide the embedded VBR functionality to give good connectivity over the IP network, as well as interoperability with the existing speech terminals. In this paper, we define that the embedded VBR coder is an audio coder that can generate variable bit-rates gracefully by the control of its application and provide scalable speech quality according to the changes of bit-rate by the structure of hierarchical bit-stream.

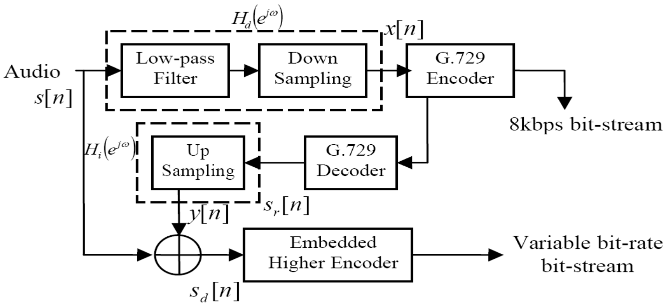

**Fig. 1.** A Block Diagram of the Embedded Audio Encoder Model

We propose a coder for the ubiquitous speech communications because its quality, media bandwidth, and interoperability can be controlled to cope with the requirements for the ubiquitous speech service. Fig. 1 and Fig. 2 show a block diagram of the embedded VBR audio encoder and decoder model. Fig. 1 shows the block diagram of the embedded VBR audio encoder accomplished in the transmitter. First, a decimator $H_d(e^{jw})$ makes the input signal down-sampled to 8 kHz

$$X(e^{jw}) = H_d(e^{jw}) \cdot S(e^{jw}), \tag{1}$$

where $S(e^{jw})$ is the input signal and $X(e^{jw})$ is the down-sampled signal. $X(e^{jw})$ is the input signal to G.729 encoder. Then, G.729 encoder constructs 8 kbit/s bit-stream for transmitting the narrowband speech from 50 Hz to 3.4 kHz. In order to compute the remaining signal which is not coped with by G.729 encoder, the decoded speech signal is generated by the G.729 decoder, then it is up-sampled to adjust a sampling rate to the original signal by an interpolator $H_i(e^{jw})$.

$$Y(e^{jw}) = H_i(e^{jw}) \cdot S_r(e^{jw}), \tag{2}$$

where is the reconstructed signal by the G.729 decoder and is the up-sampled signal. The remaining signal,, is computed by the difference between the original signal and the up-sampled reconstructed signal as

$$s_d[n] = s[n] - y[n] \tag{3}$$

In this computation, we consider the delay from G.729 coder. That is, it is considered that 5 ms look-ahead, pre- and post-processing delay times in G.729 coder and processing times needed for decimation and interpolation [8]. After computing the remaining signal, the embedded high-layer encoder is performed. It generates variable bit-rate bit-stream. Fig. 2 shows the block diagram of embedded VBR audio decoder accomplished in receiver. Transmitted bit-stream is divided into two parts, 8 kbps bit-stream and VBR bit-stream. 8 kbps bit-stream is decoded by G.729 decoder and VBR bit-stream is decoded by the embedded high-layer decoder, respectively. After up-sampling of the decoded signal by G.729 to adjust a sampling rate, two signals are merged and the complete signal $r[n]$ is reconstructed.

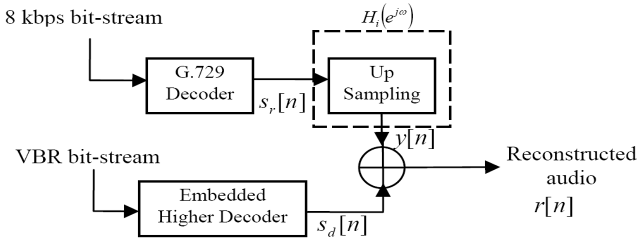

**Fig. 2.** A Block Diagram of the Embedded Audio Decoder Model

$$r[n] = x_r[n] + s_d[n], \tag{4}$$

where $x_r[n]$ is the reconstructed narrowband signal by G.729 decoder, and $s_d[n]$ is the reconstructed remaining signal by the embedded high-layer decoder. When we reconstruct the final signal, we consider the delay from G.729, the embedded high-layer coder, and an interpolator similar to computing Equation (4).

## 3 Proposed Embedded Variable Bit-Rate Audio Coder

To enhance the quality of voice for the ubiquitous applications, wideband speech codec technology was the first consideration for better quality of media source itself, since the performance of speech codec affects the quality of VoIP directly. And in order to provide robustness against the fluctuation of effective bandwidth over the ubiquitous network, we paid attention to the embedded VBR wideband speech codec described in Section 3. G.729 based embedded VBR coder(G.729EV) [9] was defined at SG16 of ITU-T. The main features of ToR of G.729EV coder are summarized in Table 1.

**Table 1.** Main Terms of Reference for G.729EV

| Terms | Requirement |
|---|---|
| Core layer | G.729 |
| Bandwidth | [300, 3400] $\sim$ [50, 7000] Hz |
| Sampling rate for input signal | 16 kHz |
| Frame size | 20 ms |
| Bit rates | 8, 12$\sim$32 kbit/sec |
| Granularity of bit-rates | 2 kbit/s |
| Algorithm delay | $<$ 60 ms |
| Complexity | $<$ 40 WMOPS |
| Quality | Not worse than G.722@56kbit/s |
| Target application | Packet Voice(VoIP) |

As shown in Table 1, the features of G.729EV provide high-quality internet telephony service for wireline and wireless networks providing two strong advantages. One is bit-level interoperability with legacy G.729 core codec, which is very popular for VoIP services in Asia, Europe, and North America. Also, since the frame size of G.729 is very short as 10ms, it gives the easier interoperability with the mobile phone. The second and main advantage of G.729EV for internet telephony over ubiquitous network is its scalability for the capacity of terminals and bandwidth. It adapts data rates according to the status of network from 8 and 12$\sim$32kbit/s with the steps of 2 kbit/s.

### 3.1 Encoder

The block diagram of the proposed encoder is given in Fig. 3. 16 kHz wideband input is low-pass filtered and then down sampled to 8 kHz. This down-sampled

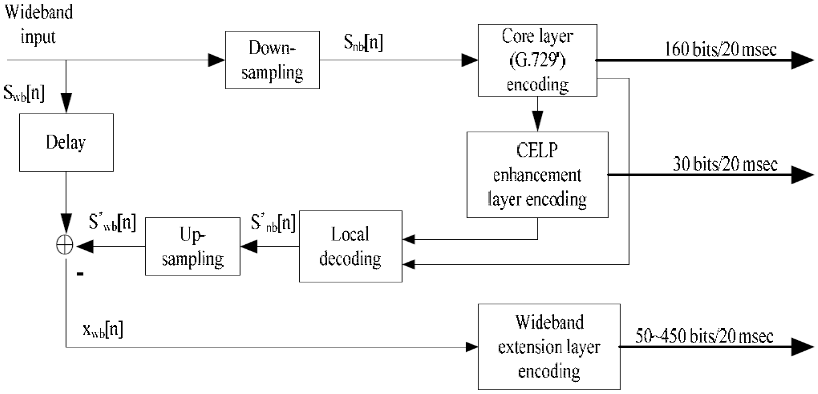

**Fig. 3.** Block Diagram of the Proposed Embedded VBR Audio Encoder

narrowband signal is encoded by the core layer and CELP enhancement layer. The core layer is based on ITU-T G.729 standard codec.

This layer is interoperable with ITU-T G.729 standard codec in the level of bitstream. The fixed codebook error signal of the core layer is processed in CELP enhancement layer in order to improve the quality of core layer. Thus the output of this layer is narrowband signal and the bit-rate is 1.5 kbit/s.

The difference signal between the delayed wideband input and up-sampled output of local decoder is processed in wideband extension layer. The difference signal is transformed using modified discrete cosine transform(MDCT). The coefficients are divided into several bands. The scale factor and normalized shape vector of each band are quantized respectively. The core layer is similar to G.729 standard codec except the LPC analysis window, pre-filtering, and post-filtering. The pre-filtering and post-filtering are suppressed. The length of the cosine part and the center location of the LPC analysis window are changed [10] and the look-ahead size is increased from 5 ms to 10 ms. The CELP enhancement layer is designed to improve the quality of core layer. In this layer, the fixed codebook error signal of core layer is represented by two algebraic pulses in every 10 ms. Signs and positions of the pulses are quantized with 15 bits. The pulses are scaled with the fixed codebook gain of core layer.

An input signal $X_{wb}$[n] of the wideband extension layer in the Fig. 3 is the difference between the delayed wideband input signal and the up-sampled version of locally decoded narrowband signal, and the signal is processed on every 20 ms frames. $X_{wb}$[n] is transformed first using MDCT. The MDCT is performed on 40 ms windowed signal with 20 ms overlap. The MDCT coefficients, X(k), is split into two typical bands, one for [0, 2.7 kHz] and the other for [2.7, 7.0 kHz]. The coefficients of the first band are quantized in MDCT domain and the coefficients of the second band are quantized on Linear Predictive Coding(LPC) residual domain. Finally, all of the quantized parameters are encoded and packed into a bitstream at the bit-packing block according to the predefined order.

### 3.2 Decoder

The decoder also comprises three layers: core layer, CELP enhancement layer and wideband extension layer as shown in Fig. 4. The operation of each layer depends on the size of the received bit stream. A frame erasure concealment algorithm is also applied in order to improve the synthesized quality in frame erasure condition. The frame erasure concealment algorithm of core layer is partly modified based on a state machine [10]. The pitch gain and fixed codebook gain is reconstructed by an attenuated version of the previous pitch gain and fixed codebook gain respectively. The attenuation coefficient depends on the state. In the case of voiced frame, the fixed codebook gain is set to zero. In wideband extension layer, an erased frame is recovered by multiplying a randomly generated shape vector by the attenuated scale factor of the previous frame.

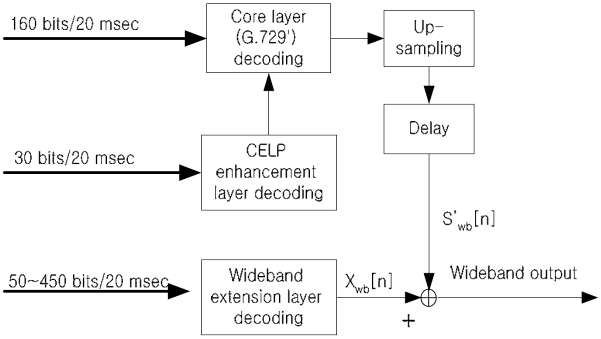

**Fig. 4.** Block Diagram of the Proposed Embedded VBR Audio Decoder

## 4 Performance Evaluation

We evaluate the performance of the proposed embedded VBR audio coder in terms of the quality, algorithmic delay, complexity, and the size of memory. The quality of proposed coder was evaluated formally by ITU-T subjective [11] and objective test methods [12]. The complexity was calculated by weighted million operations per second(WMOPS) defined in ITU-T P.191 [13].

### 4.1 Quality Evaluation

In order to evaluate the speech quality, we perform the subjective and objective tests for obtaining MOS and PESQ-WB scores which are formally defined each in the ITU-T P.800 and P.862.2 recommendations. Table 2 shows MOS score for narrowband speech, and compares its quality with the existing G.729A coder in order to evaluate the enhancement of quality.

**Table 2.** MOS Scores for narrowband speeches

| Coder | Male | Female | Mean |
|---|---|---|---|
| Direct | 4.375 | 4.229 | 4.302 |
| G.729A | 3.667 | 3.688 | 3.678 |
| Proposed@8k | 4.042 | 3.938 | 3.990 |

Table 3 shows the MOS score for the wideband speech, and compared its quality with the existing G.722 at 48 and 56 kbit/s, and G.722.2 at 8.85 kbit/s coder in order to evaluate the quality.

**Table 3.** MOS Scores for the wideband speeches

| Coder | Male | Female | Mean |
|---|---|---|---|
| Direct | 4.521 | 4.563 | 4.542 |
| G.722.2@8.85k | 4.063 | 3.917 | 3.990 |
| G.722@48k | 3.896 | 3.875 | 3.885 |
| G.722@56k | 4.146 | 4.125 | 4.135 |
| Proposed@14k | 4.313 | 4.354 | 4.333 |
| Proposed@24k | 4.313 | 4.188 | 4.250 |
| Proposed@32k | 4.458 | 4.271 | 4.365 |

Table 4 shows the mean value of PESQ-WB scores of the proposed audio coder in order to evaluate the linear quality enhancement of the embedded VBR audio coder for wideband. For the objective test, we use 5 languages; Korean, French, Japanese, German, and English.

**Table 4.** PESQ-WB Scores for the Proposed Coder

| Bit-rates(kbit/s) | PESQ-WB Score | Bit-rates(kbit/s) | PESQ-WB Score |
|---|---|---|---|
| 14 | 3.28 | 16 | 3.29 |
| 18 | 3.37 | 20 | 3.55 |
| 22 | 3.58 | 24 | 3.60 |
| 26 | 3.61 | 28 | 3.62 |
| 30 | 3.62 | 32 | 3.69 |

Table 5 also shows better quality of the proposed audio coder compared with reference coders at -16, -36 dB signal levels and noise conditions. All of the experiments involved four talkers (two males/two females), three samples per talker, and three panels of 8 listeners each (24 listeners total).

The result concludes that the proposed audio coder has 10.85% better quality than the existing G.729A coder at the same 8 kbit/s in terms of MOS score.

**Table 5.** MOS Scores at different signal levels and noise conditions

| Proposed | | Reference | |
|---|---|---|---|
| 8k (-16dB) | 3.760 | G.729A@8k (-16dB) | 3.688 |
| 8k (-36dB) | 3.896 | G.729A@8k (-36dB) | 3.500 |
| 32k (-16dB) | 4.188 | G.722A@56k (-16dB) | 4.344 |
| 32k(-36dB) | 4.292 | G.722A@56k (-16dB) | 3.458 |
| 8k (Music) | 4.240 | G.729A@8k (Music) | 4.219 |
| 8k (Office) | 4.198 | G.729A@8k (Office) | 4.083 |
| 8k (Babble) | 4.625 | G.729A@8k (Babble) | 4.354 |
| 32k (Music) | 4.396 | G.722A@56k (Music) | 3.969 |
| 32k (Office) | 3.396 | G.722A@56k (Office) | 4.219 |
| 32k (Babble) | 4.458 | G.722A@56k (Babble) | 4.188 |

At 14kbit/s mode of the proposed coder which is the minimal bit-rate of wideband, it shows 10.86% quality enhancement than G.722.2 at 8.85kbit/s. And at 32kbit/s mode of the proposed coder, it shows 10.86∼11.24% quality enhancement than G.722 at 48 and 56 kbit/s.

### 4.2 Algorithm Delay

The algorithm delay of the proposed audio coder is 40.75ms, which comprises 20 ms framing delay, which is the same value of frame size of the coder, 10ms look-ahead, 10ms MDCT overlapping window, and 0.75ms up/down sampling delay.

### 4.3 Complexity and Memory

To calculate the complexity for the implementation of the proposed coder, we use the WMOPS which are defined in ITU-T P.191 recommendations. The complexity and the size of memory are summarized in the Table 6 and Table 7 respectively. In table 6, the complexity is evaluated in the worst case. The total values are given by the sum of the three layers and other functions such as resampling. The overall complexity of the proposed codec is about 37.85 WMOPS.

**Table 6.** The Worst Case Computational Complexity(WMOPS) of the Proposed Coder

| Components | Encoder | Decoder |
|---|---|---|
| Core layer | 11.683 | 2.612 |
| CELP enhancement layer | 5.707 | 0.042 |
| Wideband enhancement layer | 9.692 | 4.265 |
| Other functions | 2.519 | 1.372 |
| Total | 29.601 | 8.291 |

**Table 7.** Memory Requirement of the Proposed Coder(Word)

| Memory Types | Encoder | Decoder | Total |
|---|---|---|---|
| PROM | 3,704 | 2,943 | 6,647 |
| DROM | 22,865 | | 22,865 |
| DRAM | 4,295 | 3,897 | 8,192 |
| Total | 37,704 | | |

The DROM takes into account all the constant tables. Same tables are used in both encoder and decoder. Thus the DROM in table 6 is the summation of the encoder and decoder. The DRAM corresponds to the memory of all the static variables and worst case of the dynamic RAM usage.

## 5 Conclusion

In this paper, we have proposed an embedded VBR audio coder in order to provide the fittest quality of service and better connectivity of service for the speech communications over the ubiquitous network environment. After dividing input signal into narrowband and wideband, it performs coding procedure in each part hierarchically and the bit-rate for providing the fittest quality between the ubiquitous end points is determined dynamically according to the channel conditions and terminal capacities. This embedded VBR architecture of the proposed audio coder provides the fittest speech quality of service, better connectivity of service, and excellent service completion ratio among the various ubiquitous end points. Therefore, the proposed audio coder is useful for high-quality speech communications such as voice over IP, conversational e-learning, audio conferencing, remote monitoring, conversational internet games, and other multimedia services over the ubiquitous network. For the interoperability with legacy voice terminal, the proposed audio coder has the compatibility with existing G.729 coder. Moreover, it uses G.729 enhancement coder to improve quality of the narrowband signal and to provide the basis of better quality of the higher band. This higher band, which covers all bandwidth of human speech, provides better quality compared with the voice quality of analog telephones. The merit of the interoperability of the proposed coder enables wider reuse of the existing voice over IP systems such as G.729/G.729a terminals and gateways. As a result, the proposed coder has a reasonable performance compared with the existing wideband audio coder by the subjective and objective evaluation measures of speech quality.

**Acknowledgments.** Authors acknowledge the contribution of Jongmo Sung, Mi Suk Lee, and Hyun-Woo Kim in the development of the embedded VBR coder. The work was supported by the Ministry of Information and Communication, Republic of Korea.

## References

1. Do Young Kim, Mi Suk Lee, H.W.J.H.K.K.: Scalable speech and audio coding technologies for wireless network. In: Proc. of KICS. Volume 22., Seoul, KICS, KICS (2005) 1397–1407
2. G.729: Coding of speech at 8kbps using conjugate-structure algebraic-code-excited linear-prediction (cs-celp). In: ITU-T Recommendation, Geneva, ITU, ITU-T (1996)
3. G.729A: G.729 annex a: Reduced complexity 8 kbit/s cs-acelp speech codec. In: ITU-T Recommendation, Geneva, ITU, ITU-T (1996)
4. G.729B: G.729 annex b: A silence compression scheme for g.729 optimized for terminals conforming to recommendation v.70. In: ITU-T Recommendation, Geneva, ITU, ITU-T (1996)
5. G.711: Pulse coded modulation(pcm) of voice frequencies. In: ITU-T Recommendation, Geneva, ITU, ITU-T (1988)
6. G.722: 7 khz audio coding within 64 kbit/s. In: ITU-T Recommendation, Geneva, ITU, ITU-T (1988)
7. G.722.2: Wideband coding of speech at around 16kbit/s using adaptive multi-rate wideband (amr-wb). In: ITU-T Recommendation, Geneva, ITU, ITU-T (2002)
8. G. H. Lee, Y. H. Lee, H.K.K.D.Y.K., Lee, M.S.: A scalable audio coder for high-quality speech and audio services. In: Proc. of the 9th Western Pacific Acoustics Conference, Seoul (2006) 178–185
9. ITU-T: Q10/16 meeting report, Geneva, ITU, ITU-T (2004)
10. ITU-T: High-level description of etri candidate codec for g.729ev, Geneva, ITU, ITU-T (2005)
11. P.800: Methods for subjective determination of transmission quality, Geneva, ITU, ITU-T (1996)
12. P.862.2: Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs, Geneva, ITU, ITU-T (2005)
13. P.191: Software tools for speech and audio coding, Geneva, ITU, ITU-T (1993)