A Scalable Audio Coder for High-quality Speech and Audio Services*

이길호, 이영한, 김홍국(GIST), 김도영, 이미숙(ETRI)

<차 례>

<Abstract>

A Scalable Audio Coder for High-quality Speech and Audio Services

Gil Ho Lee, Young Han Lee, Hong Kook Kim, Do Young Kim, Mi Suk Lee

In this paper, we propose a scalable audio coder, which has a variable bandwidth from the narrowband speech bandwidth to the audio bandwidth and also has a bit-rate from 8 to 320 kbits/s, in order to cope with the quality of service(QoS) according to the network load. First of all, the proposed scalable coder splits bandwidth of the input audio into narrowband up to around 4 kHz and above. Next, the narrowband signals are compressed by a speech coding method compatible to an existing standard speech coder such as G.729, and the other signals whose bandwidth is above the narrowband are compressed on the basis of a psychoacoustic model. It is shown from the objective quality tests using the signal-to-noise ratio(SNR) and the perceptual evaluation of audio quality(PEAQ) that the proposed scalable audio coder provides a comparable quality to the MPEG-1 Layer III (MP3) audio coder.

* Keywords: Scalable audio coding, Speech and audio coding, G.729, MP3, PEAQ

* This research was supported by the Ministry of Information and Communication(MIC), Korea, under the Information Technology Research Center(ITRC) support program supervised by the Institute for Information Technology Advancement(IITA-2006-C1090-0603-0017) at Gwangju Institute of Science and Technology(GIST).

1. Introduction

In the current state-of-the-art digital communications networks, two kinds of speech coders are used for the voice communication services. One is a narrowband speech coder that compresses speech signals whose bandwidth is defined as 3.4 kHz. On the other hand, the advancement of digital communications technologies enable us to deal with wideband speech signals whose bandwidth is defined as around 7 kHz. The existing wideband speech coders provide higher quality speech communication than the narrowband speech coders. However, since these two kinds of coders are optimized based on the assumption that the input signals are speech even if they can deal with music signals, the decoded quality by the coders are much different from that of the natural sound. One of the major reasons of the difference is that the bandwidth has been limited up to 7 kHz while the bandwidth of the natural sound(audio) should be extended up to 20 kHz. Another reason is that audio signals must be considered as a different way that speech signals are mainly assumed to be periodic in time and frequency depending on the pitch period of speech.

On the one hand, many research works have been introduced for high-quality audio coding. However, there are redundant bits compared to speech coders if they are used for speech communications because audio coders focus on maximizing the perceived quality. Depending on the network conditions such as the type of input signals, the network congestion, and so on, scalable audio coding can trade off the audio service quality and a bit-rate. One of the scalable approaches is to design an audio coder which split audio signals into constant width sub-bands in the frequency domain and compressed each sub-band signal with different number of bits according to the specified bit-rate, where the bit-rate is changed from 30.2 to 61.6kbits/s [1]. A bit-stream scalable audio coder that split audio signals into low frequency components and high frequency components was proposed in [2], where a hybrid warped linear predictive coding(WLPC)-wavelet representation is used to encode the low frequency components but the high frequency components are encoded using an LPC noise model. As another approach, scalable coders based on the analysis of audio source characteristics have been developed. Ramprashad proposed the multimode transform predictive coder(MTPC) [3]. The MTPC operates as one of three primary modes such as speech, music, and transitional modes, where the mode is decided by using present and past pitch delays, long-term prediction gains, and LPC filter parameters. In [4], the authors proposed a scalable coder by combining the well-known speech coding algorithms, where G.729 and the time domain bandwidth extension(TDBWE) [5] were

used as a baseline coder and an enhancement layer, respectively.

In this paper, we propose a scalable audio coder for high-quality speech and audio communication services according to the type of the input signals and the network load. The proposed audio coder has a variable bandwidth from the narrowband speech bandwidth to the audio bandwidth, and can operate at a bit-rate varying from 8 to 320 kbits/s. Moreover, for the backward compatibility with a standard speech coder, the proposed audio coder has the standard speech coder as a narrowband core speech coder.

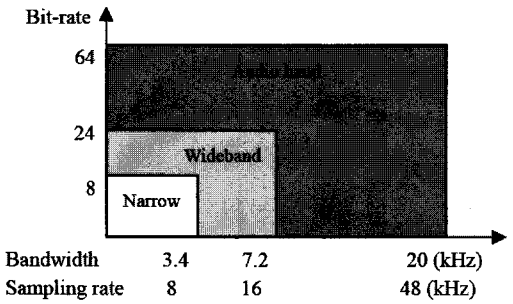
Following this introduction, in Section 2, we briefly review the coding techniques that are the basis of the proposed scalable audio coder, which include narrowband speech coding, wideband speech coding, and audio coding using a psychoacoustic model. In Section 3, we propose a scalable audio coder. In Section 4, we evaluate the performance of the proposed scalable audio coder operating at 128 kbits/s in terms of the signal-to-noise ratio and the objective difference grade. Finally, we present our conclusions in Section 5.

2. Speech/audio Coder and Psychoacoustic Model

The audible frequency of the human hearing is known to be ranged from 20 Hz to 20 kHz [6]. <Figure 1> shows this audible frequency range which can be divided by three parts in a view of source coding. We define these bands as narrowband, wideband, and audio band. Narrowband speech coders have been developed to compress the human speech whose bandwidth is restricted below 4 kHz. One of the popularly used speech coders is the 8 kbits/s CS-ACELP which is the ITU-T Recommendation G.729 [7]. According to the advancement of network technology and internet services, most of users demand high quality speech services. Recently, G.722.2 was standardized for a wideband speech service over W-CDMA [8]. The wideband coder is optimized for providing good quality for speech signals and even can treat music signals as an input, but its quality needs to be improved compared to that of the audio coders such as MPEG-1, 2, and 4 audio coders. This is because it is usual for a wideband coder usually not to cope with audio band signals.

On the other hand, audio coders can compress audio band signals with high quality, where audio band signals are sampled at a rate of up to 48 kHz. The main difference between audio coders and narrowband/wideband speech coders is that audio coders process perceptible signals in the human hearing system, while speech coders

are dominated by using model-based approaches such as linear predictive coding and sinusoidal coding. In a view of the human hearing, audio coding adopts a psychoacoustic model that is based on the masking properties of hearing and utilized for designing a quantizer [9]. Most of the audio coders such as MP3, AAC, and AC3 follow this approach even with different kinds of psychoacoustic models [10].

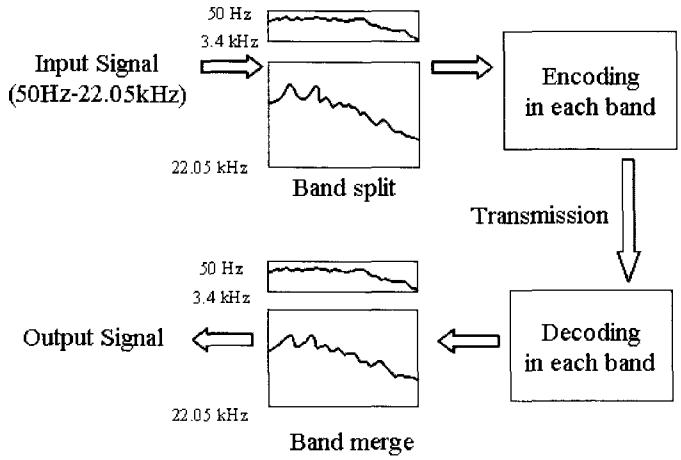


<Figure 1> Speech/audio bandwidth sampling rate and available bit-rate

3. Proposed Audio Coding Method

In the previous section, we briefly reviewed narrowband speech coding, wideband speech coding, and audio coding. In order to get high quality, an audio coding scheme must be adopted. However, our goal is to develop an audio coder with a special need that a newly designed audio coder can be backward compatible with the existing narrowband speech codec. In this paper, we propose a scalable audio coder consisting of the G.729 speech coder as a narrowband core coder and a psychoacoustic audio coder as an enhancement layer.

<Figure 2> shows a procedure of the proposed scalable audio coding method. The frequency of the input signals is located from 50 Hz to 22.05 kHz. This range corresponds to the audible frequency of the human hearing system. At first, the spectrum of the input signals is divided into two bands. In order to provide backward compatibility with the narrowband speech codec, narrowband signals from 50 Hz to 3.4 kHz are encoded by a narrowband speech coder.



<Figure 2> A procedure of the proposed scalable audio coding

Wideband and audio band signals above 3.4 kHz are encoded by a psychoacoustic coder. According to the channel condition, the proposed scalable audio coder decides its operating mode. In other words, if the bit rate is restricted into 8 kbits/s by the network, it only encodes narrowband of the input signals by using the core coder, G.729. On the other hand, the full band of the input signals by using G.729 with a psychoacoustic enhancement layer is encoded with a bit rate of up to 320 kbits/s. The decoder of the proposed scalable audio coder generates decoded signals of each band separately on the basis of the bit-streams, and then combines to make the reconstructed full band audio signals.

<Figure 3> shows the block diagram of the proposed scalable audio encoder. In order to obtain the narrowband signals from the input signals, a decimator is first applied to the input signals. Decimator is implemented by using 256-point FFT as low-pass filtering. In other words, the narrowband spectrum is represented as

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

where $S(e^{jw})$ is the input signal and $H_d(e^{jw})$ is the frequency response of the decimator to 8 kHz sampling frequency. The decimated signal, $x[n]$, is processed by the G.729 encoder and represented by the 8 kbits/s bit-stream. In order to compute the remaining signal that is not covered by the G.729 encoder, the 8 kbits/s bit-stream is decoded by the G.729 decoder. After that, the decoded signal is up-sampled to adjust

a sampling rate to the original signal by an interpolator $H_i(e^{jw})$. Interpolator is also implemented by using 256-point FFT as low-pass filtering. As a result, we have the spectrum, $Y(e^{jw})$, that is already modeled by the G.729 speech coder as

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

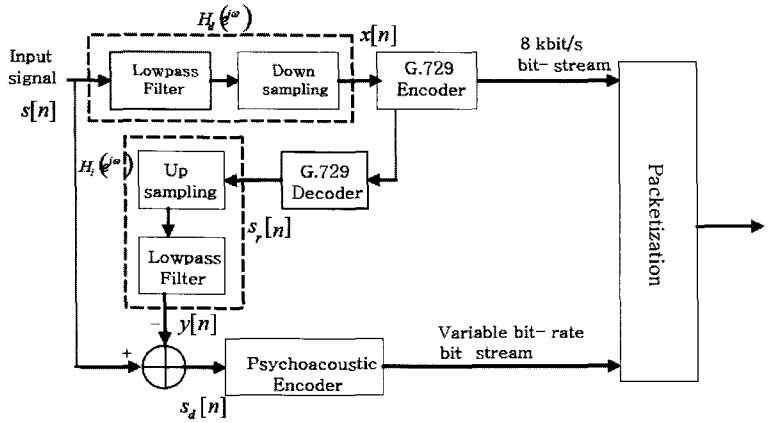
where $S_r(e^{jw})$ is the narrowband signal decoded by G.729. The remaining signal, $s_d[n]$, is computed by the difference between the original signal and the up-sampled reconstructed signal as $s_d[n] = s[n] - y[n]$.

When computing the difference of $y[n]$ from $s[n]$, the processing delays should be considered. They include a delay for decimation from 44.1 kHz to 8 kHz and a delay for interpolation from 8 kHz to 44.1 kHz. In addition, G.729 encoding makes 5 msec look-ahead, pre- and post-processing delays. Next, the psychoacoustic encoder such as MPEG-1 Layer III(MP3), Advanced Audio Coding(AAC) is applied to $s_d[n]$. Finally, the bit-streams from the G.729 encoder and the psychoacoustic encoder are packed and transmitted to the decoder of the proposed scalable audio coder.

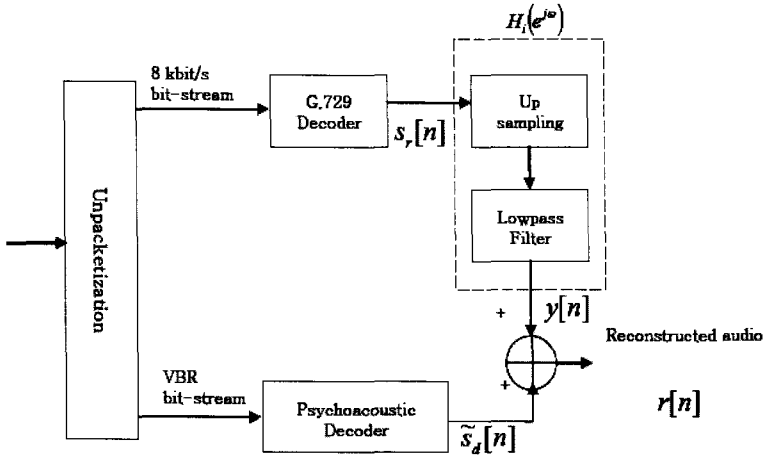
<Figure 4> shows the block diagram of the proposed scalable audio decoder. The transmitted bit-stream is split into two parts: the 8 kbits/s bit-stream and the remaining variable bit-rate(VBR) bit-stream. The 8 kbits/s bit-stream is decoded by the G.729 decoder and the VBR bit-stream is decoded from the psychoacoustic decoder. After up-sampling the decoded signal from 8 kHz to 44.1 kHz sampling frequency, the audio signal $r[n]$ is reconstructed by adding the decoded signal by the psychoacoustic decoder, $\widetilde{s_d}[n]$, and the up-sampled signal from the decoded signal by the G.729 decoder, $y[n]$, as

$$r[n] = y[n] + \widetilde{s_d}[n] \tag{3}$$

Similarly to the encoding, we consider the delay when merging the two signals, where the delays are caused by the G.729 decoding delay, the delay for the interpolation, and the delay for the psychoacoustic decoding.



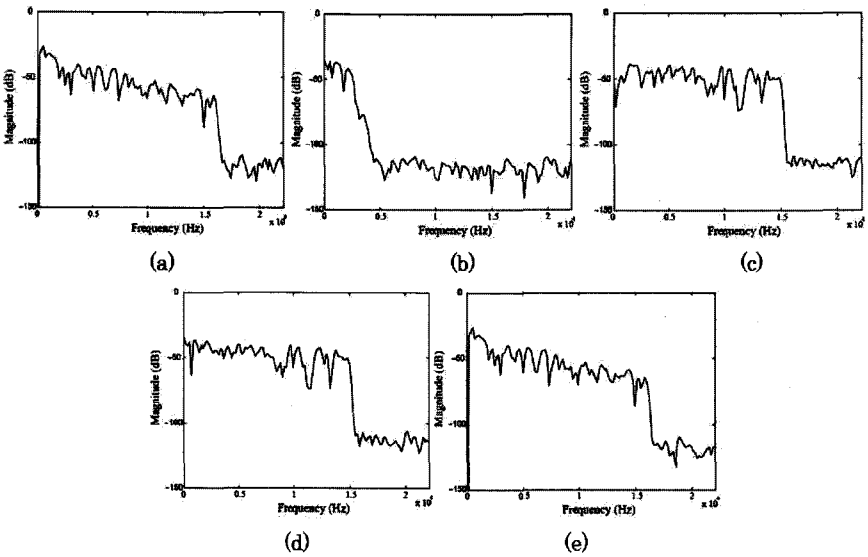
<Figure 3> A block diagram of the proposed scalable audio encoder



<Figure 4> A block diagram of the proposed scalable audio decoder

<Figure 5> shows the magnitude spectra of audio signals in each step of the proposed scalable audio coder. <Figures 5(a)-(d)> show the magnitude spectrum of the input audio, $S(e^{jw})$, the narrowband magnitude spectrum decoded by G.729, $S(e^{jw})$, the audio band magnitude spectrum decoded by the psychoacoustic decoder, $\tilde{S}(e^{jw})$, and the magnitude spectrum of the reconstructed audio, $R(e^{jw})$, respectively. As a reference, the magnitude spectrum of the reconstructed audio by a MP3 is shown in <Figure 5(e)>, where the MP3 coder was operated with a bit rate of 128 kbits/s. Although the spectrum of the reconstructed signal, $R(e^{jw})$, is a little different from that of the input audio, but it will be shown in the next section that the listeners could not notice any

perceptual difference between the decoded audio signals by MP3 and by the proposed scalable audio coder.



<Figure 5> Comparison of spectra obtained from (a) the original signal sampled at a rate of 44.1 kHz, (b) the decoded signal by G.729, (c) the decoded signal by the psychoacoustic decoder part (d) the reconstructed audio signal by the proposed scalable coder and (e) the coded signal by MP3

4. Performance Evaluation

In this section, we evaluated the performance of the proposed scalable audio coder by measuring two objective measures. One was the signal-to-noise ratio(SNR) and the other was the objective difference grade(ODG). We chose six types of audio sources such as male voice, female voice, male pop, female pop, crossover, and symphony. Each type consisted of four audio files, which resulted in 24 audio files as a total; each audio file has which have 5 minutes length. First of all, each audio file was prepared by sampling at a rate of 44.1 kHz and quantizing samples with a 16-bit uniform quantizer. The proposed coder provided bit-rates from 8 to 328 kbits/s with a step of 8 kbits/s since the bit-rate of the baseline coder was fixed as 8 kbits/s but the enhancement layer had variable bit-rates from 8 to 320 kbits/s with a step of 8 kbits/s. Here, we modified MP3 quantization levels so that MP3 could operate at

bit-rates of 8, 16, and 24 kbits/s. The framework employed in the proposed coder can be applied to other audio coders by replacing MP3 with one of the audio coders including AAC and AC-3. This can result in providing the different range of bit-rates to the proposed coder. To compare the quality of the proposed scalable audio coder with that of the existing standard audio coder, we processed the audio files by MP3 with a bit rate of 128 kbits/s. Also, the proposed coder has a bit rate of 128 kbits/s by assigning 8 kbits/s to G.729 and 120 kbits/s to the psychoacoustic coder.

4.1. Segmental SNR

As an objective measure, the segmental SNR(SegSNR) was computed by the following equation.

$$SegSNR = \frac{1}{M}\sum_{k=0}^{M-1}\left(10\log_{10}\frac{\sum_{n=0}^{N-1}s^2[kN+n]}{\sum_{n=0}^{N-1}(s[kN+n]-r[kN+n])^2}\right) \text{ (dB)} \quad (4)$$

where M is the number of segments, N is the number of samples in a segment, $s[n]$ is the original audio signal, and $r[n]$ is the reconstructed audio signal. <Table 1> shows the comparison of SegSNR between the coded audio signals by the MP3 coder and the proposed scalable audio coder for 6 audio types. From <Table 1>, it was shown that the performance of the proposed scalable audio coder was similar to that of MP3 for all audio types.

4.2. Perceptual evaluation of audio quality

For another objective sound quality measure, we used the perceptual evaluation of audio quality(PEAQ) [11]. PEAQ has been widely used to compare a listener's subjective judgment of state-of-the-art perceptual audio coders in an objective way. For a given pair of audio files, a 5-grade objective difference grade(ODG) defined by the ITU-R Recommendation BS.562-3 [12] was obtained from the PEAQ test. <Table 2> shows the 5-grade impairment scale defined by the ITU-R Recommendation BS.562-3 [12], where each listener gives a score depending on how much the processed auido is

likely to be impaired compared to the reference audio. The description of each point is also shown in the table. <Table 3> shows ODG scores of MP3 and the proposed coder for 6 audio types. It was shown from the table that the sound quality of the proposed coder was similar to that of MP3.

<Table 1> Comparison of segmental SNR of MP3 and the proposed coder for different audio types

Audio Type	SegSNR(dB)	
	MP3 (128 kbits/s)	Proposed coder (128 kbits/s)
Male voice	34.94	34.42
Female voice	39.04	38.84
Pop(male)	31.40	31.06
Pop(female)	29.13	28.61
Crossover	35.58	34.39
Symphony	41.15	40.71

<Table 2> ITU-R 5-grade impairment scale

Impairment Description	ODG
Imperceptible	0.0
Perceptible, but not annoying	-1.0
Slightly annoying	-2.0
Annoying	-3.0
Very annoying	-4.0

<Table 3> Comparison of objective difference grade(ODG) of MP3 and the proposed coder for different audio types

Audio Type	Objective difference grade	
	MP3 (128 kbits/s)	Proposed coder (128 kbits/s)
Male voice	-0.16	-0.22
Female voice	-0.21	-0.21
Pop(male)	-0.05	-0.10
Pop(female)	-0.06	-0.15
Crossover	-0.08	-0.10
Symphony	-0.15	-0.11

5. Conclusion

In this paper, we proposed a scalable audio coder for high quality simultaneous speech and audio services. The proposed coder was designed to have a variable bandwidth from the narrowband speech bandwidth to the audio bandwidth and have a variable bit-rate from 8 to 328 kbits/s with a step of 8 kbits/s, in order to cope with the quality of service(QoS) according to the network load. The proposed coder performed coding separately for each band after splitting input signals into two parts such as a narrowband and a higher band. Therefore, this structure could support backward compatibility with the existing narrowband speech coder, especially G.729, and use a psychoacoustic model for representing the higher band. The performance of the proposed coder was compared with that of the MP3 coder by measuring the signal-to-noise ratio and the perceptual evaluation of audio quality score. As a result, it was found out that the proposed scalable audio coder had a comparable performance to the MP3 coder with a bit-rate of 128 kbits/s, supporting a different range of bandwidths from speech to audio band.

References

[1] O. V. D. Vrecken, L. Hubaut, F. Coulon, "A new subband perceptual audio coder using CELP", *Proc. ICASSP*, pp. 3661-3664, May 1998.

[2] D. Ning, M. Deriche, "A bitstream scalable audio coder using a hybrid WLPC-wavelet representation", *Proc. ICASSP*, pp. 417-420 Apr. 2003.

[3] S. A. Ramprashad, "A multimode transform predictive coder(MTPC) for speech and audio", *Proc. IEEE Workshop on Speech Coding*, pp. 10-12, June 1999.

[4] M. Meuleneire, H. Taddei, O. Zelicourt, D. Pastor, P. Jax, "A CELP-wavelet scalable wideband speech coder", *Proc. ICASSP*, pp. 697-700, May 2006.

[5] P. Jax, B. Geiser, S. Schandl, H. Taddei, P. Vary, "An embedded scalable wideband codec based on the GSM EFR codec", *Proc. ICASSP*, pp. 5-8, May 2006.

[6] E. Zwicker, H. Fastl, *Psychoacoustics: Facts and models*, Springer, 1990.

[7] ITU-T Recommendation G.729, *Coding of speech at 8 kbits/s using conjugate-structure algebraic-code-excited linear prediction(CS-ACELP)*, Mar. 1996.

[8] ITU-T Recommendation G.722.2, *Wideband coding of speech at around 16 kbits/s using adaptive multi-rate wideband(AMR-WB)*, Jan. 2002.

[9] ISO/IEC JTC/SC29/WG11 MPEG IS11172-3, *Information technology - coding of moving picture and associated audio, Part 3: Audio*, 1992.

[10] R. N. J. Velhuis, "Bit rates in audio source coding", *IEEE Transactions on Selected Areas in Comm.*, Vol. 10, No. 1, pp. 86-96, Jan. 1992.

[11] ITU-R Recommendation BS.1387-1, *Method for objective measurements of perceived audio quality*, Nov. 2001.
[12] ITU-R Recommendation BS.562-3, *Subjective assessment of sound quality*, Oct. 2004.

접수일자: 2007년 2월 14일
게재결정: 2007년 3월 22일

▶ 이길호(Gil Ho Lee)
주소 : Maetan 3 Dong Yeongtong Gu, Suwon 443-742, Korea
소속 : Software Engineering Team, Software Laboratories, CTO, Samsung Electronics
전화 : +82-31-277-7871
E-mail : adelio.lee@samsung.com

▶ 이영한(Young Han Lee)
주소: 1 Oryong-dong, Buk-gu, Gwangju 500-712, Korea
소속: Dept. of Information and Communications, GIST
전화: +82-62-970-3121
E-mail: cpumaker@gist.ac.kr

▶ 김홍국(Hong Kook Kim) : Corresponding Author
주소: 1 Oryong-dong, Buk-gu, Gwangju 500-712, Korea
소속: Dept. of Information and Communications, GIST
전화: +82-62-970-2228
E-mail: hongkook@gist.ac.kr

▶ 김도영(Do Young Kim)
주소: 161 Gajeong-dong, Yuseong-gu, Daejeon 305-350, Korea
소속: BcN Research Division, ETRI
전화: +82-42-860-5180
E-mail: dyk@etri.re.kr

▶ 이미숙(Mi Suk Lee)
주소: 161 Gajeong-dong, Yuseong-gu, Daejeon 305-350, Korea
소속: BcN Research Division, ETRI
전화: +82-42-860-6148
E-mail: lms@etri.re.kr