# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| Title | Bandwidth Extension with Hybrid Signal Extrapolation for Audio Coding' |
| Authors | Chatree Budsabathon, Akinori Nishihara |
| Citation | IEICE Trans.Fundamentals., Vol. E90-A, No. 8, pp. 1564-1569 |
| Pub. date | 2007, 8 |
| URL | http://search.ieice.org/ |
| Copyright | (c) 2007 Institute of Electronics, Information and Communication Engineers |

IEICE TRANS. FUNDAMENTALS, VOL.E90–A, NO.8 AUGUST 2007

PAPER Special Section on Papers Selected from the 21st Symposium on Signal Processing

# Bandwidth Extension with Hybrid Signal Extrapolation for Audio Coding

**Chatree BUDSABATHON**[†a)], ***Student Member and*** **Akinori NISHIHARA**[†b)], ***Fellow***

**SUMMARY** In this paper, we propose a blind method using hybrid signal extrapolation at the decoder to regenerate lost high-frequency components which are removed by encoders. At first, a decoded signal spectral resolution is enhanced by time domain linear predictive extrapolation and then the cut off frequency of each frame is estimated to avoid the spectrum gap between the end of original low frequency spectrum and the beginning of reconstructed high frequency spectrum. By utilizing a correlation between the high frequency spectrum and low frequency spectrum, the low frequency spectrum component is employed to reconstruct the high frequency spectrum component by frequency domain linear predictive extrapolation. Experimental results show an effective improvement of the proposed method in terms of SNR and human listening test results. The proposed method can be used to reconstruct the lost high frequency component to improve the perceptual quality of audio independent of the compression method.

***key words:*** *perceptual audio coding, bandwidth extension, linear predictive extrapolation*

## 1. Introduction

High frequency components are removed when we convert analog audio signals to digital or when we convert digital audio signals to other digital audio signals with lower sampling rate or lower bit rate by a perceptual audio coding. In currently existing low bit rate high quality audio compression, the sampling rate and audio bandwidth are always restricted corresponding to the storages and communication bandwidth. When the quantization bit is limited, most perceptual audio compressions do not allocate bits to the high frequency subband and put all available bits to the lower frequency subbands which are more perceptive for human auditory systems. For example, a well-known MPEG-1 Layer3 (MP3) restricts the bandwidth to less than 16 kHz and 12 kHz by default at bit rate 128 kbps and 64 kbps, respectively.

Missing high frequency components decrease the quality of audio signals such as localization, ambient information and bright nature of audio. To improve the quality of those audio signals, the audio bandwidth extension for regenerating the lost high frequency components have been proposed. Most of the bandwidth extension methods have been proposed for narrowband signals and speech [1], [2]. For general audio signals, the state of art "spectral band replication (SBR)" [3] is a method to reduce the bit rate or improve the audio quality by mapping the low-frequency part of an audio signal coded at low bit-rate to the missing high-frequency region then the audio bit stream energy envelope is shaped by a side information embedded with the data. It has been combined to MP3 and called MP3-Pro and also combined with MPEG-4. The newer method is accurate spectral replacement [4]. In that method, a signal is at first normalized using a smooth spectral envelope model then the flattened signal is segmented into sinusoids and noise part. The synthesis of sinusoids and stationary noise is combined and de-normalized using the smooth spectral envelope model. Those methods need to modify the encoder for calculating the necessary information of the high frequency components, so the decoder can regenerate the high frequency components correctly. Audio signals that are encoded or decoded by any perceptual audio coding without that function will not get any advantage from that method. Therefore a reconstruction method in which the decoder does not require any information about the high frequency band of the original signal is sometimes preferred. A Blind method using non-linear device was proposed in [5]. In that method, a pre-bandpass filter is used to pick the lower frequencies from the input audio signal and to remove the unwanted high frequency component. Then the high frequency components around the original signal are generated using a non-linear device such as a half wave rectifier. The post-bandpass filter is used to filter only the high-frequency components from the generated signal. The high frequency band signal level is adjusted by gain G and then finally the upper band is summed with the delayed low band original signal. The other method was proposed in [6], where the envelope of the high frequency component is calculated by a linear estimation of the known signal spectra in the frequency domain. Then a certain width of spectrum selected from the low frequency signal is copied to the high frequency using the defined envelope.

This paper presents a blind bandwidth extension method to reconstruct the lost high frequency components

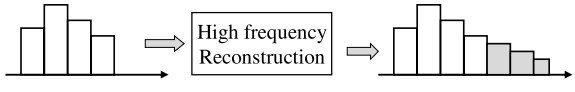


**Fig. 1** Bandwidth extension for audio signal.

Manuscript received December 21, 2006.
Manuscript revised March 17, 2007.
Final manuscript received April 17, 2007.

[†]The authors are with the Department Communications and Integrated Systems, Tokyo Institute of Technology, Tokyo, 152-8552 Japan.
a) E-mail: chatree@nh.cradle.titech.ac.jp
b) E-mail: aki@cradle.titech.ac.jp

DOI: 10.1093/ietfec/e90–a.8.1564

which are close to the original spectrum for general audio signals as shown in Fig. 1. A decoded signal spectral resolution is enhanced by time domain linear predictive extrapolation and then the cut off frequency of each frame is estimated to avoid the spectrum gap between the low frequency part of the input signal. By utilizing a correlation between the high frequency spectrum and low frequency spectrum, low frequency spectrum components are used to reconstruct high frequency spectrum components by frequency domain linear predictive extrapolation. The method presented in this paper does not need additional information from either encoders or decoders so most of the encoded bandwidth limited audio can be improved in the perceptual quality by our proposed method.

The organization of this paper is as follows. Section 2 presents a hybrid signal extrapolation method starting with a linear predictive extrapolation, a method to detect the cut off frequency and high frequency reconstruction process. The experimental results show the improvement of the proposed method in Sect. 3. Finally, the conclusion is given in Sect. 4.

## 2. Hybrid Signal Extrapolation Method

It is well known that when the signal is transformed from time domain to frequency domain by Fast Fourier Transform (FFT), the higher spectral resolution can be achieved by increasing the FFT frame size. If the signal has a transient in that frame, however, the dynamic property is lost and Inverse Fast Fourier Transform (IFFT) would not recover the original signal. So the frame size cannot be made long. The main procedure in our proposed method for each data frame is:

**Step:1** Enhance the signal spectral resolution by extrapolating the future and past samples using a linear predictive extrapolation in the time domain.
**Step:2** Detect the cut off frequency dynamically to avoid the spectrum gap between the input signal.
**Step:3** Estimate high frequency spectra through over the cut off frequency in the frequency domain by forward linear predictive extrapolation of the low frequency envelope spectra.

This method is based on the hypothesis that both the low frequency components and the high frequency components are outcomes of the same physical process (e.g. the high frequency components are harmonics of the low frequency component from vocal and/or musical instruments) [7]. The block diagram of our method is shown in Fig. 2. At first, the spectrum resolution of decoded PCM signal $x(n)$ ($0 < n \leq N_1$, $N_1$ samples/frame) is enhanced by forward and backward time domain linear predictive extrapolation and then the extrapolated signal ($N_2$ samples/frame) is transformed into frequency domain by windowing the signal using a Hamming window and FFT. The cut off frequency $k_c$ is detected dynamically by averaging the spectrum power and comparing with the threshold value. The envelope of frequency component higher than $k_c$ is predicted by predictive extrapolation. The low frequency spectrum $|X'_L(e^{j\omega})|$ is used to replicate the high frequency detail spectrum. The reconstructed high frequency components combined with the original low frequency band $|X'_{L+H}(e^{j\omega})|$ is transformed back to the time domain via IFFT. The signal is de-windowed by dividing the signal with the Hamming window and then is truncated by cutting the forward and backward samples, and finally, is processed by an overlap-add method to generate the output $y(n)$.

### 2.1 Linear Predictive Extrapolation of Audio Signal

The wideband audio signal in nature is non stationary. When we process a frame of signal in the frequency domain, the transformed spectral resolution depends on the number of samples within one frame and the signal should stay relatively stationary within the frame. This leads to a tradeoff between spectral and temporal resolutions. The time linear predictive extrapolation is a method to expand the signal in one frame both forward and backward direction by modelling the unknown signal using the known data. The extrapolated signal in time domain can increase the resolution in frequency domain. The conventional method for mathematical modelling of speech signals is the linear predictive coding (LPC) where the $n^{th}$ known signal sample $x_n$ is approximated as a linear combination of $p$ previous samples and residual error computed using a finite impulse response (FIR) filter as

$$x(n) = -\sum_{i=1}^{p} a_i x(n-i) + e(n), \tag{1}$$

where $a_i$ are the prediction coefficients and $p$ is the model

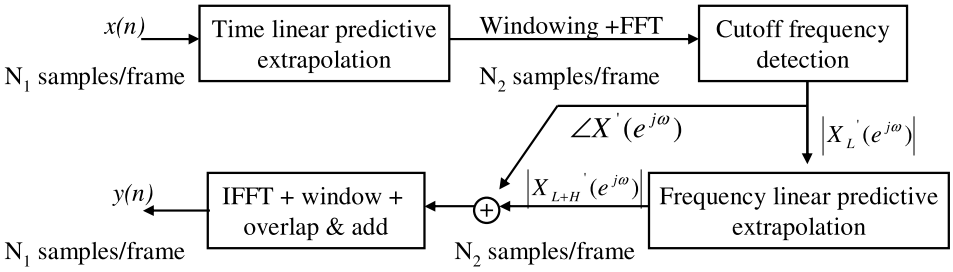


**Fig. 2** Block diagram of proposed system.

order and $e(n)$ is a residual error. The coefficient $a_i$ are calculated by solving the linear equations using Levinson-Durbin recursion to minimize the square error of $e(n)$. The $p$ is set to 8–12 for speech and 30–100 for audio signals [9].

The short stationary audio signal can be approximated by the combination of $L$ signal spectrum as

$$x(n) = \sum_{i}^{L} A_i(n\Delta t)\cos(2\pi f_i n\Delta t + \phi_i) + \epsilon(n\Delta t) \tag{2}$$

where $f_i \geq 0$, $A_i$ and $\phi_i$ are the amplitude and the phase of each frequency $f_i$, $\Delta t$ is the sampling interval, and $\epsilon$ is noise. In linear predictive extrapolation, the signal is divided into a short frame length $N$ then this signal is modelled using Eq. (1). The unknown samples $x_{N+1}, x_{N+2}, \ldots$ are calculated by the known signal samples $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]$. The forward extrapolation and the backward extrapolation are defined by modifying the LP equation as

$$x^f(n) = -\sum_{i=1}^{p} h_i^f x_{(n-i)}, \qquad n > N \tag{3}$$

$$x^b(n) = -\sum_{i=1}^{p} h_i^b x_{(n+i)}, \qquad n < 1 \tag{4}$$

respectively, where $x^f(n)$, $x^b(n)$ are the forward and backward extrapolated samples. $h_i^f$, $h_i^b$ are the forward and backward extrapolation model coefficients. The Burg method [9] is well known for calculating the AR coefficients and then converting them into impulse response coefficients $h_i$. It needs around 1000 impulse coefficient $h_i$ to model the frame size around 2000 samples of audio signal. By using too short impulse responses to extrapolate the given signal frame, only the strongest frequencies in the signal are extrapolated while the lower one or random noise can not be properly extrapolated. Therefore we use the low order (30–100) for the propose to suppress the random noise. The procedure for the extrapolation of new $W$ samples is:

1. Calculate the impulse response coefficients $h_1$, $h_2, \ldots, h_p$ via Burg method.
2. Initialize the filter with $p$ past known samples before the part to be extrapolated.
3. To obtain $W$ extrapolated samples, feed a zero vector of length $W$ as an input to the filter.

The original signal and the linear predictive extrapolated signal in time domain are shown in Fig. 3, and the frequency response of the both signal are shown in Fig. 4. We can see that the peak spectra are not changed and the spectrum resolution is enhanced as shown in Fig. 4.

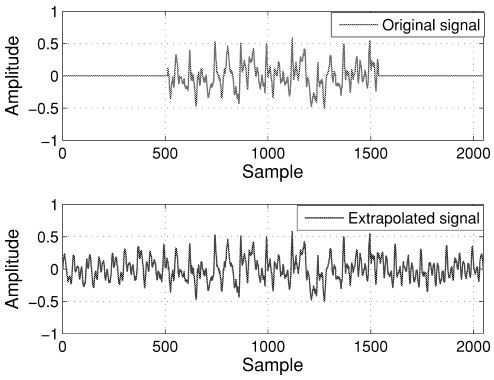


**Fig. 3** Linear predictive extrapolation in time domain.

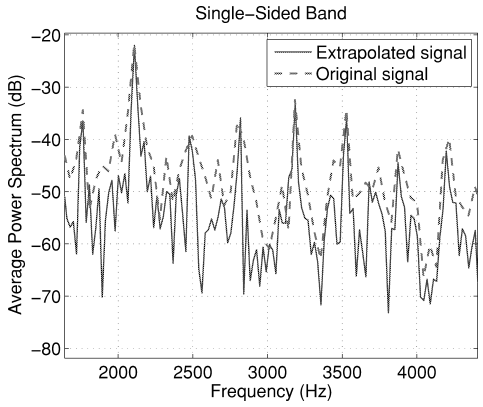


**Fig. 4** Frequency response of time domain extrapolated signal.

## 2.2 Cut Off Frequency Detection

In perceptual audio encoder such as MP3, the bit allocation divides the data bit to each subband according to the information of audio signal, therefore, the high frequency subband where zero bit is allocated is not always the same. If the beginning of the extension frequency is higher than the cut off frequency, there will be a gab between them, which causes a poor reconstruction. To estimate the cut off frequency $k_c$, at first we calculate the average spectrum $|\bar{X}|$ of $P$ frames as

$$|\bar{X}| = \sum_{i=1}^{P} |X(i)|/P, \tag{5}$$

where $P$ is chosen between 3 frames to 20 frames. If $P$ is set as a large value, the average spectrum is smooth and easy to determine the cut off frequency but the cut off frequency of each frame may be slightly differ from the determined average cut off frequency. Next, finding the frequency bin $f_t$ whose average spectrum, $|X(f_t)|$ is larger than the threshold value $\alpha$ as

$$f_t = T \qquad \text{if} \qquad |X(T)| - |\bar{X}| > \alpha, \tag{6}$$

where $\alpha$ is chosen between $-40$ dB and $-60$ dB and the $X(T)$ is calculated from the largest frequency bin $T$ then step down the frequency bin value. The cut off frequency $k_c$ is determined as the frequency which is lower than the threshold frequency in the specified range $\beta$ to ensure the continuity of reconstructed frequency as illustrated in Fig. 5.

$$k_c = f_t - \beta, \tag{7}$$

where $\beta$ is chosen between 200 Hz and 1000 Hz.

### 2.3 High Frequency Reconstruction

The method proposed by [6] used the linear line to approximate the envelope in dB scale of high frequency component. However, the characteristic of audio spectrum is not always linearly decreasing. In our method, after the time domain linear extrapolated audio signal is transformed into frequency domain, the high frequency signals are reconstructed in logarithm scale of magnitude and linear scale of frequency using linear predictive extrapolation in the frequency domain. Let $|X(k)|$ be a spectrum at frequency bin $k$. The envelope of the high frequency through over the cut off frequency $k_c$ is estimated by the linear predictive extrapolation of the spectra at frequency lower than the cut off frequency as shown in Fig 6. The spectrum beyond $k_c$ $(X[k], k > k_c)$ can be approximated using the same procedure as that used in time domain extrapolation. Each estimated high frequency coefficient $X[k]$ is a weighted combination of the $p$ lower frequency coefficients as

$$|X[k]| = \sum_{i=1}^{p} H_i |X[k-i]| \qquad k > k_c, \tag{8}$$

where $H_i$ are the forward frequency domain extrapolation model coefficients with minimum mean square error. The algorithm for calculate the impulse response $H_i$ is the same as in time domain.

Since human can not detect phase distortion of audio signal, the phase of the compressed audio signal is combined with the generated high frequency spectrum $|X[k]|$ and then transformed back to the time domain by IFFT and overlap-add method as

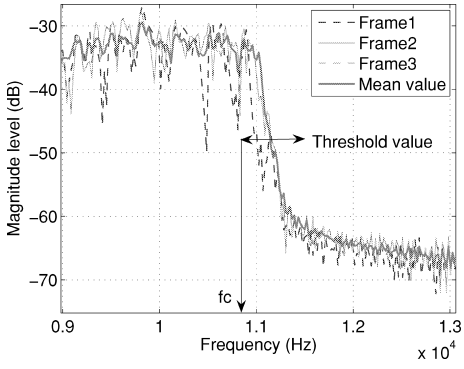


**Fig. 5** Dynamically cut off frequency detection.

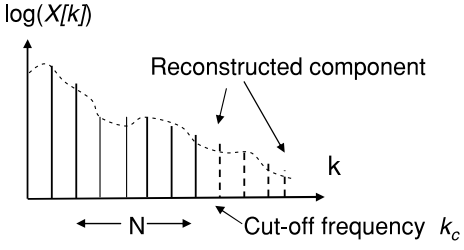


**Fig. 6** High frequency component reconstruction.

$$y(n) = \mathrm{IFFT}\left\{|X(k)|e^{j\theta_k}\right\}, \tag{9}$$

where $\theta_k$ is the phase of the original signal.

## 3. Experimental Simulation Results

Test audio signal samples are prepared by compressing original CD quality (1.44 Mbps) of wideband audio signals of various types such as violin, piano, orchestra, pop music etc, into bit rates at 128 kbps and 96 kbps with sampling rate at 44.1 kHz (stereo). A general commercial powerpack lame MP3 codec [10] is used to make the low bit rate test samples. The operating frame size is 20 ms or 1024 samples. The order range for time domain and frequency domain linear extrapolation is about 10 to 50. The spectrum of original audio signal and compressed audio signal are shown in Figs. 7 and 8, respectively. We can see that the bit rate is not enough so that the signal spectrum over 14 kHz is almost set to zero. By our proposed algorithm the high frequency spectrum is reconstructed as shown in Fig. 9.

The spectrogram of the original CD and signal coded by MP3 are shown in Figs. 10 and 11. Since the coding

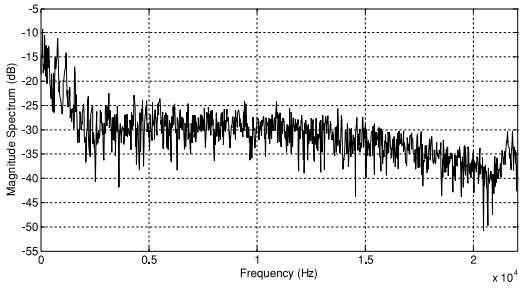


**Fig. 7** Spectrum of the original audio signal.

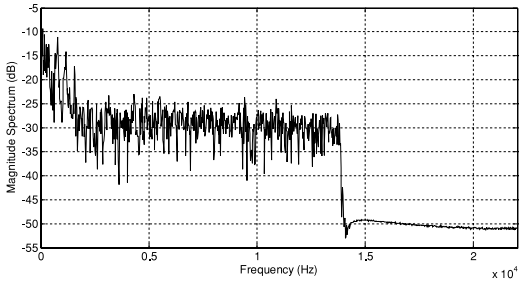


**Fig. 8** Spectrum of the compressed audio signal.

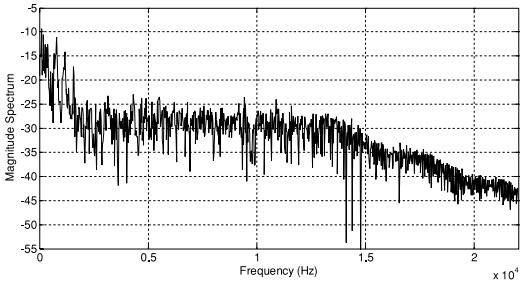


**Fig. 9** Spectrum of the audio with proposed method.

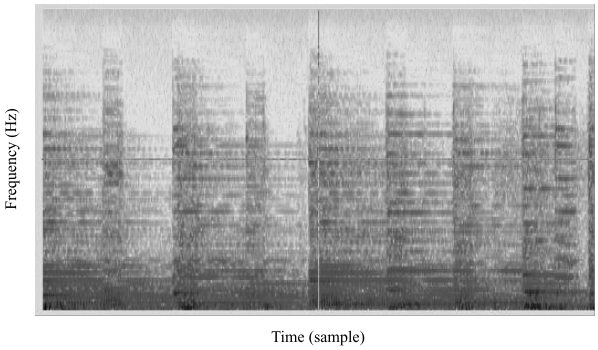


**Fig. 10** Spectrogram of original music signal.

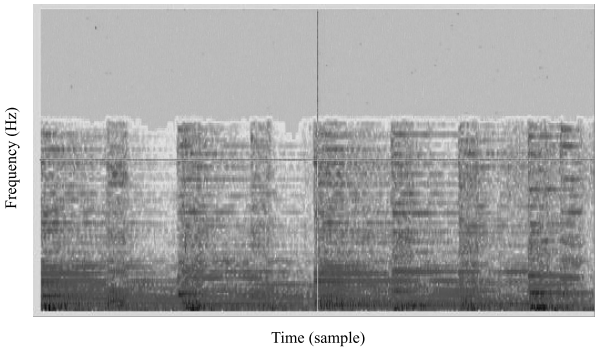


**Fig. 11** Spectrogram of coded signal by MP3.

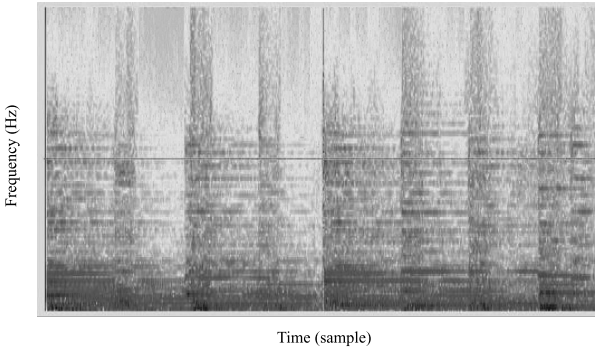


**Fig. 12** Spectrogram of signal after high frequency reconstruction.

bit rate is very low, high frequency bands are suppressed because no bits are allocated there. Figure 12 clearly shows that the proposed system can regenerate the high frequency component that is very similar to the original signal so that the signal quality is enhanced and sound more feeling and brighter compared with the MP3 signal.

The objective measure in terms of SNR is used to evaluate the signal quality enhancement after bandwidth extension. The definition of SNR is

$$SNR = 20\log_{10}\frac{\sum_{i=1}^{N}|X_{or}(i)|}{\sum_{i=1}^{N}(|X_{or}(i)| - |X_{out}(i)|)}, \quad (10)$$

**Table 1** SNR results.

| Test sample | SNR (dB) | | | | |
|---|---|---|---|---|---|
| | harpsichord | opera | orchestra | violin | vocal |
| MP3 | 14.64 | 20.07 | 17.54 | 16.79 | 14.59 |
| Proposed | 15.02 | 21.16 | 18.01 | 17.44 | 15.25 |

**Table 2** Results of human listening tests.

| Audio Signal | Average probability of listener thought that signal with proposed method is better than signal with conventional method |
|---|---|
| Pop music | 1.0 |
| Percussion | 0.75 |

where $X_{or}$ is the original CD signal spectrum and $X_{out}$ is the spectrum under test. Table 1 shows the SNR of signal after processed by our proposed method and the MP3 audio signal. Our proposed method can improve the SNR a few decibels. Simulation results show that the bandwidth extension with time domain and frequency domain extrapolation gave a little higher SNR than only the frequency domain extrapolation or MP3.

The subjective tests were also carried out to confirm the effective improvement by the proposed method. Four university staff including experts in listening to high quality audio signals are trained to understand the perceptual quality degradation due to the lost high frequency components. The subjects are screened so that they can properly distinguish CD quality and MP3 sounds. The test sample signals are prepared by compressing the stereo CD signals to 96 kbps. The sampling rate is 44100 Hz and the audio length is 10 seconds. We used a high quality headphone which can give an output frequency range from 5 Hz to 30 kHz. The comparison test between our method and the conventional method proposed in [6] is performed. First the subjects listen to the original CD quality audio signal. Next the signals coded by the conventional method [6] and the coded signals processed by our proposed method were presented randomly. Therefore, listeners did not know which is the proposed or the conventional one. Then they were asked to select the sample with quality close to the original CD signal and gave a comment about the signal quality. The listeners can listen to the original one and the test samples as many times as they want. The results are averaged and given in Table 2, which shows the superiority of the proposed method over the conventional method. In some audio sample coded at 96 kbps, the reconstructed high frequency component such as "sh" in vocal sounds similar to the quantization noise. This can be improved by lower the prediction order to reduce the effect of non-voiced signal.

## 4. Conclusion

This paper presents a method to reconstruct the lost high-frequency component at the decoder by hybrid extrapolation for low bit rate audio coding. The signal spectral resolution is increased by time linear predictive extrapolation and the high frequency spectra are reconstructed in frequency do-

main by linear predictive extrapolation of the low frequency envelope spectra. The experimental results show that the high frequency components can be reconstructed and the audio sound is better than unprocessed audio sound. This method can be used with other kind of low bit rate high quality compressed audio signal, independent of the encoding algorithm.

**References**

[1] C. Avendaro, H. Hermansky, and E.A. Wan, "Beyond Nyquist: Towards to recovery of broad-bandwidth speech from narrow-bandwidth speech," Proc. Eurospeech, pp.165–168, 1995.
[2] J. Valin and R. Lefebvre, "Bandwidth extension of narrowband speech for low bit-rate wideband coding," IEEE Speech Coding Workshop, pp.130–132, Sept. 2000.
[3] M. Dietz, L. Liljeryd, K. Kjörling, and O. Kunz, "Spectral band replication, a novel approach in audio coding," AES 112th Convention, Munich, Germany, May 2002.
[4] A. Ferreira and D. Sinha, "Accurate spectral replacement," 118th AES Convention, 2005.
[5] E. Larsen, R.M. Aarts, and M. Danessis, "Efficient high-frequency bandwidth extension of music and speech," AES 112th Convention, Munich, Germany, May 2002.
[6] C.M. Liu, W.C. Lee, and H.W. Hsu, "High frequency reconstruction for band-limited audio signals," Proc. DAFX-03, Sept. 2003.
[7] P. Jax and P. Vary, "An upper bound on the quality of artificial bandwidth extension of narrow band speech signals," Proc. ICASSP. vol.1, pp.237–240, Orlando, FL, May 2002.
[8] S. Godsill, P. Rayner, and O. Cappe, Digital Audio Restoration, Springer-Verlag, London, 1998.
[9] I. Kauppinen and K. Roth, "Audio signal extrapolation-theory and applications," Proc. DAFX-02, pp.105–110, Sept. 2002.
[10] The LAME Project, website http://www.mp3dev.org/

**Chatree Budsabathon** was born in Bangkok, Thailand. He received the B.Eng. degree (with first class honors) from Chulalongkorn University, Bangkok, Thailand, in 2000. Since 2000, he has received the Japanese Government Scholarship and has continued his education in Japan as the research student. He received the M.Eng. and Dr. Eng. degrees in Communications and Integrated Systems Engineering from Tokyo Instituted of Technology, Tokyo, Japan, in 2003 and 2007 respectively. His main research is in digital signal processing techniques for multimedia and communication.

**Akinori Nishihara** received the B.E., M.E. and Dr. Eng. degrees in electronics from Tokyo Institute of Technology in 1973, 1975 and 1978, respectively. Since 1978 he has been with Tokyo Institute of Technology, where he is Professor of the Center for Research and Development of Educational Technology. His main research interests are in signal processing, and its application to educational technology. He served as an Associate Editor of the IEICE Trans. Fundamentals from 1990 to 1994, an Associate Editor of the IEEE Transactions on Circuits and Systems II from 1995 to 1997, and Editor-in-Chief of Trans. IEICE Part A from 1998 to 2000. He served as an ExCom member of IEEE Region 10 (Asia Pacific Region), as Student Activities Committee Chair, Treasurer, and Educational Activities Committee Chair. He now serves as a member of IEICE Strategic Planning Committee, a member of IEEE EAB Committee of Global Accreditation Activities, Chair of IEEE Circuits and Systems Society Japan Chapter, and adviser of IEEE Japan Council Women in Engineering Affinity Group. He received the IEICE Best Paper Award in 1999 and the IEEE Third Millennium Medal in 2000. Dr. Nishihara is a Fellow of IEEE, a member of EURASIP, ECS, JSET and IEEK.