THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.22, No.3E 2003. 9. pp. 101~109

Multi Mode Harmonic Transform Coding for Speech and Music

Jonghark Kim*, Jae-Hyun Shin*, Insung Lee*
*Dept. of Radio Engineering, Chungbuk National University

(Received September 18 2002; revised February 7 2003; accepted February 17 2003)

Abstract

A multi-mode harmonic transform coding (MMHTC) for speech and music signals is proposed. Its structure is organized as a linear prediction model with an input of harmonic and transform-based excitation. The proposed coder also utilizes harmonic prediction and an improved quantizer of excitation signal. To efficiently quantize the excitation of music signals, the modulated lapped transform (MLT) is introduced. In other words, the coder combines both the time domain (linear prediction) and the frequency domain technique to achieve the best perceptual quality. The proposed coder showed better speech quality than that of the 8 kbps QCELP coder at a bit-rate of 4 kbps.

Keywords: *Speech coding, Harmonic coding, CELP, Audio coding*

I. Introduction

Recently, there has been increasing interests in coding of speech and audio signals for several applications such as audio/video teleconferencing, wireless multimedia, wideband telephony over packet networks, and Internet applications. These applications usually require the modeling method of mixture signals such as speech and audio[1-4]. Compression algorithms designed specifically for speech or audio signals, such as music, have been successfully deployed in application areas such as telecommunications, digital broadcasting, and storage. In many instances, however, the algorithms were designed for the particular input signal or application, thus they have not been met quality expectations for the broader class of input signals. Until recently, algorithms[5,6] designed for both speech and other, more diverse, audio signals, have not received considerable attentions. Recent progress in this area, however, has shown that increased quality levels at low bit rates could only be achieved at the expense of higher algorithmic delay, or complexity, or a compromised quality for more diverse signals.

Specially, a harmonic coding scheme showed a good quality at demanding complexity and delay at low rate speech coder, because the coding method uses only the simple extracting structure for excitation signal by adopting a frequency domain method different from algebraic code excited linear prediction (ACELP). For the music signal, however, the harmonic signal analysis for speech model, only based on one fundamental frequency, is difficult to produce satisfactory quality. Although the structure such as peak continuation is added, the complexity is very high.

In this paper, a hybrid scheme is presented which produces satisfactory results for both speech and music signal. The proposed algorithm utilizes time-domain linear prediction (LP) and pitch prediction analysis to determine the reconstructed signal. However, instead of using the computationally demanding analysis-by-synthesis tech

Corresponding author: Jonghark Kim (easthawk@hanmail.net)
Dept. of Radio Engineering, Chungbuk National University, San-48, Gaesin-dong, Cheongju, Korea

niques to determine the innovative excitation, the perceptually weighted signal with removed filtering and pitch correlations, better known as the target signal, is transformed and encoded in the frequency domain; it is then decoded and inverse transformed to extract the time-domain innovative excitation. The coding methods for music and speech include fast Fourier transform (FFT) and modulated lapped transform (MLT). The basic scheme of the proposed coder is introduced in section 2 and speech/music discriminator is explained in section 3. In detail, harmonic noise coding and MLT excitation coding is explained in the section 4 and section 5. Then, the quantization including the bit allocation is explained in section 6 and section 7.

II. Overall Structure of Mutimode Harmonic Transform Coding

The encoder and decoder of the proposed coder are presented in Figure 1 and Figure 2. As shown in Figure 1, the encoder consists of three encoding modes for excitation signal: a harmonic excitation mode, a CELP without pitch analysis mode and a transform coded excitation mode. The U/V detector and Speech/Music discriminator is used to determine operating mode. The preprocessing and linear prediction (LP) analysis are common for all of the three modes. The harmonic excitation mode operates on 20 ms frames in voiced frame. The CELP without pitch analysis mode operates for unvoiced frames with a length of 10 ms sub-frame and the transform coded excitation mode operates for music frames with a length of 20 ms frame. The harmonic coding is based on sinusoidal wave model. The main features of harmonic excitation coding include fast Fouriour transform (FFT), spectrum pick peaking process, vector quantizer (VQ) of harmonic spectral magnitude parameters for analysis part, and inverse vector quantizer (VQ) of harmonic spectral magnitude parameters, linear phase generation, inverse fast Fouriour transform (IFFT), overlap/add synthesis for synthesis part.

In the proposed speech coder (MMHTC), the harmonic coding defines a spectral magnitude estimation error using basis function which is obtained by DFT of hamming

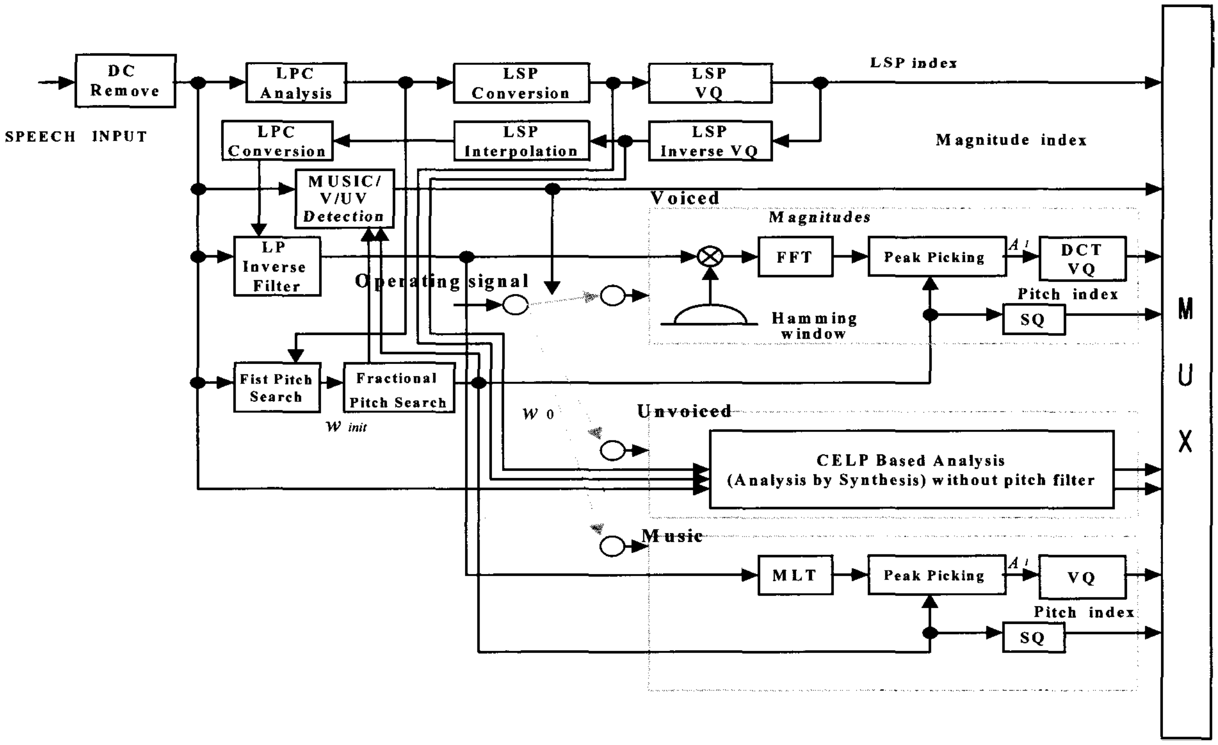

Figure 1. Block diagram of the proposed MMHTC encoder.

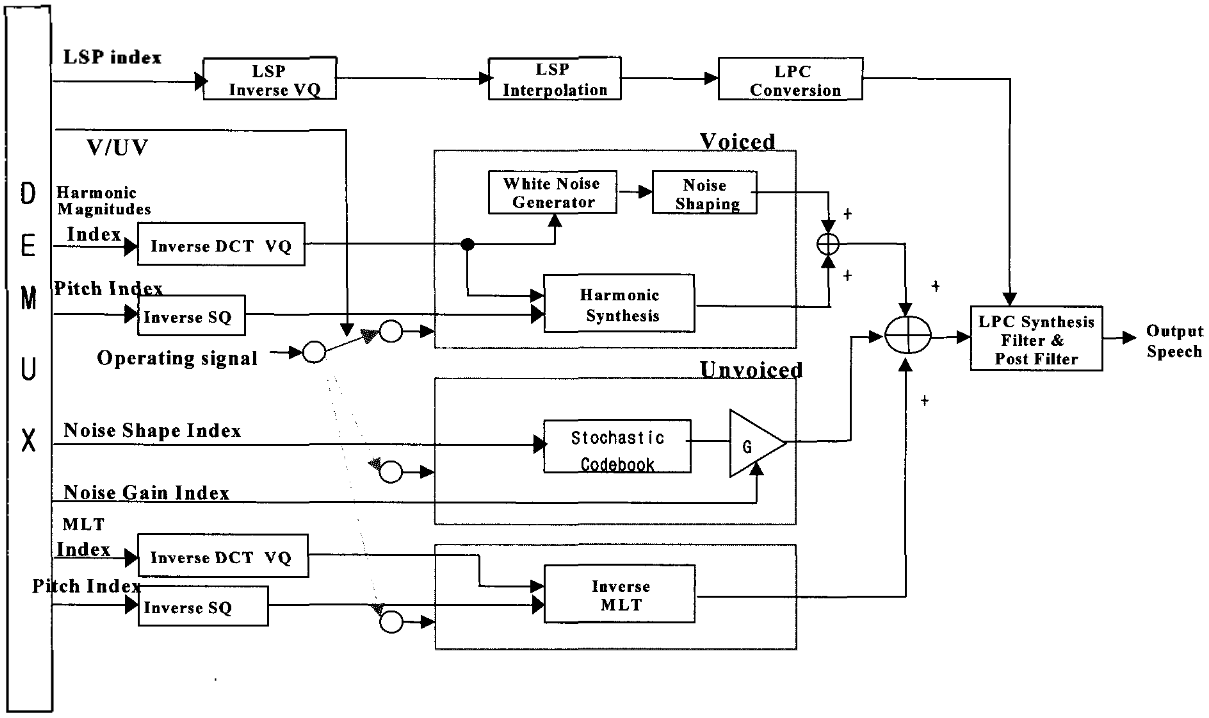

Figure 2. Block diagram of the proposed MMHTC decoder.

window and find the magnitudes to minimize the estimation error[7]. A new noise coding called by cepstrum-lpc noise coding is introduced. The cepstrum-lpc noise coding separate noise components from mixture signal and extract the spectral envelope information of unvoiced signal. Also, harmonic synthesis method is implemented by inverse fast Fouriour transform (IFFT) to reduce complexity. Since the unvoiced signals have a characteristic of random signal, the contribution of pitch information is not much. Thus, the CELP coding without the additive bits for pitch period may be a proper method. The transform-coded excitation consists of computing the target signal for the 20 ms frame followed by modulated lapped transform (MLT) and adaptive position VQ in frequency domain. The spectral envelope by LPC is utilized to quantize MLT peak coefficients.

III. Speech/Music Classification

Speech waveforms have very regular patterns with high amplitude quasi-periodic voiced segment and have higher variation than music signal in terms of energy gain and spectral line. In other word, the harmonic content of music signals is more stationary than speech signal. To discriminate speech and music, the variance of frame energies and the pitch strengths are computed by

$$V_s(n) = \sqrt{\sum_{k=0}^{N} (E_s(n-k) - \bar{E}_s)^2} \tag{1}$$

$$V_p(n) = \sqrt{\sum_{k=0}^{N} (S_p(n-k) - \bar{S}_p)^2} \tag{2}$$

where, $E_e(n)$ is the frame energy of n-th frame, $\bar{E}_e$ is mean of the frame energy. $S_p(n)$ is pitch strength of nth frame, $\bar{S}_p$ is mean of pitch strength. The pitch strength is the first peak value of normalized autocorrelation. The classification use the characteristic that speech data has higher variances than music data in the same range of mean value of pitch strength. The discrimination process is similar to MPEG 4 Music/Speech Classification Tool[7].

IV. Harmonic Noise Coding for Speech

The harmonic coding requires fundamental frequency during an encoding stage. We use a pitch obtained from closed loop pitch search whose criterion is minimizing the error between the synthesized spectrum and original spectrum. The spectral magnitudes of the synthesized spectrum are regarded as harmonic components in the harmonic plus noise coding. While, the noise components can be also estimated from valley part of the spectral magnitudes separated by the fundamental frequency and cepstrum information. These both components are coded using two other schemes according to the mode discriminated by the energy level of the noise components.

4.1. Harmonic Plus Noise Structure

The speech signal is represented by a convolution of excitation signal and vocal track impulse response. Specifically, the excitation signal consists of quasi-periodic and aperiodic part, where the quasi-periodic part means glottal pulse train of pitch period and the aperiodic part means noise-like signal due to airflow from lungs or radiation from lips.

The left part of pitch period in the quefrency domain (cepstrum) is a component due to vocal track response that have smooth spectral envelope. While, the right quefrency part of the pitch period is an excitation component[8]. Specially, the values around peak at pitch period are classified into harmonic components because harmonics are consists of multiple times of fundamental frequency. To determine noise region, the cepstrum around the peak at pitch period is liftered and is converted into log magnitude spectrum. Then, the noise component region is defined into negative part of the log magnitude spectrum[8].

To extract spectral envelope information of noise components, we applied LPC analysis to time samples given by inverse fast Fouriour transform (IFFT) of the spectrum in the noise component region. It is a similar process to estimate spectral envelop of formants for original speech signal. The LPC parameters for noise components are transmitted to decoder.

The LPC coefficients of all-pole model are converted into line spectrum pairs (LSP) parameters and quantized by full search vector quantizer with four dimensions. The synthesis processing of the decoder is simply implemented as linear prediction (LP) filtering with an input of white Gaussian noise without the phase matching between each frame. The order of all-pole model is 4 and 4, 2 bits for LSP parameters and gain parameter are assigned, respectively.

4.2. Synthesis of Harmonic Components

To guarantee continuiety in consecutive frames, the harmonic parameters must be interpolated using previous

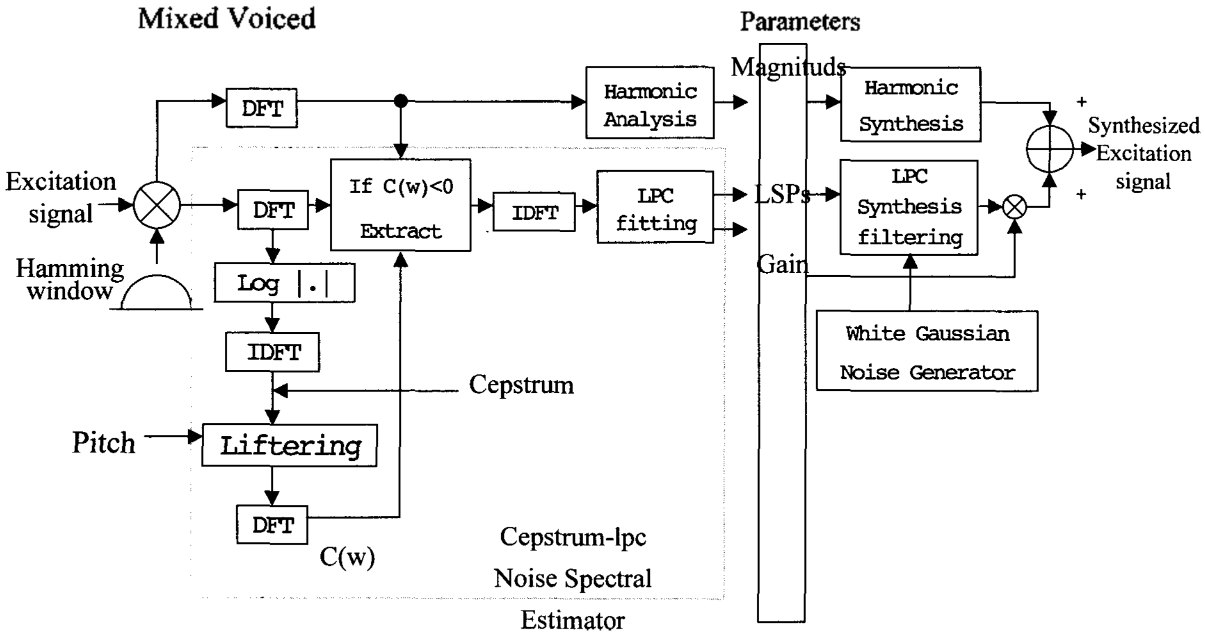

Figure 3. Block diagram of mixed coding.

parameter values. The simple linear interpolation are utilized for magnitude parameters. The phase parameters are synthesized by considering the wrapping and simutaneous matching to fundamental frequency continuation and initial phase between frames. First, we assume that the temporal fundamental frequency varies linearly. The synthesized speech is represented by[9]

$$s^k(n) = \sum_{l=1}^{L^k} A^k(l)\cos(\Delta\Phi^k(n)l + \Phi^k(l)), \qquad n = 1\cdots N \tag{3}$$

where,

$$A^k(l) = \alpha(n)A_q^k(l) + (1-\alpha(n))A_q^{k-1}(l) \tag{4}$$

$$\Delta\Phi^k(n) = \sum_{m=0}^{n}(\alpha(n)w_0^k + (1-\alpha(n))w_0^{k-1}) \tag{5}$$

where, $\alpha(n) = n/N$ is the linear incresing function, $S^k(n)$ is reconstructed signal, $A^k(l)$ is the spectral magnitudes, $\Delta\Phi^k(n)$ is the temporal phase term, $\Phi^k(n)$ is the initial phase term. k is the current frame number and l is the number of harmonic sequence, N is the frame size. Here, the initial phase $\Phi^k(n)$ is given by

$$\Phi^k(l) = \frac{Nl}{2}(w_0^{k-1} + w_0^{k-2}) + \Phi^{k-1}(l) \tag{6}$$

to satisfy $s^{k-1}(N) = s^k(0)$, it guaranties continuiety between the previous frame and the current frame. This synthesis method requires high computational complexity. The fast method can be derived by defining the basic waveform $w(m,k)$. The definition is given by

$$w(m,r) = \sum_{l=0}^{B} A_q^r(l)\cos(m\frac{2\pi}{B}l + \Phi^k(l)) \quad \text{if } l > L \text{ then } A_q^r(l) = 0 \tag{7}$$

Then, the synthesized speech is represented by

$$s^k(n) = \alpha(n)w(\frac{B}{2\pi}\Delta\Phi^k(n),k) + (1-\alpha(n))w(\frac{B}{2\pi}\Delta\Phi^k(n),k-1) \tag{8}$$

whrere B is the DFT block size. If the B is the square number of 2, then the basic waveform $w(m,k)$ can be given by

$$\begin{aligned} w(m,r) &= \mathrm{Re}\{\sum_{l=0}^{B}\left(A_q^r(l)\cos(\Phi^k(l)) + jA_q^r(l)\sin(\Phi^k(l))\right)e^{-m\frac{2\pi}{B}l}\} \\ &= \mathrm{Re}\{\mathrm{IFFT}\{A_q^r(jw)\angle\Phi^k(jw)\}\} \end{aligned} \tag{9}$$

The complexity of harmonic synthesis process is reduced significantly by (9).

4.3. Quantization of Harmonic Parameters

We assume a pitch period of speech ranging from 66 hz to 660 hz[14]. Then, the harmonic magnitudes have a variable dimension ranging from 6 to 60. To quantize a variable dimension vector, the encoder converts the variable-dimension vector into a fixed-dimension vector using a linear function. Similarly, the decoder uses the inverse linear function to convert the decoded fixed-dimension vector into a variable-dimension vector. This general approach was proposed in[10] and is called non-square transform or non-square transform VQ.

We tested a few structures to efficiently quantize the magnitudes with a variable dimension; a) the first method quantizes only the front 20 magnitudes similar to mixed excitation linear prediction (MELP) coder[11], b) the second method quantizes the vectors splitting into two same dimension groups using discrete cosine transform (DCT) transformation, c) the third method quantizes DCT coefficients after splitting the magnitudes into two groups with the same dimension, d) the fourth method quantizes using two stages VQ with the same dimension. The spectral distortion is measured by[10]

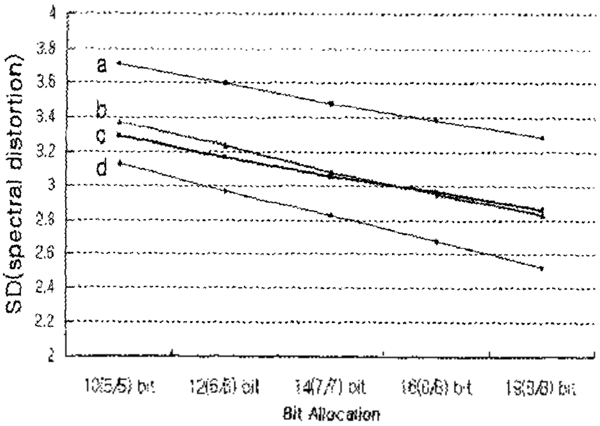

Figure 4. Spectral distortion for the proposed vector quantizers of harmonic magnitudes.

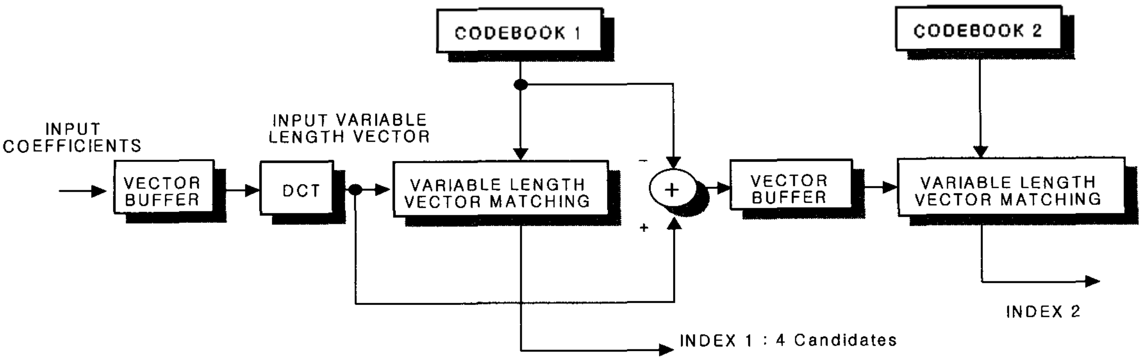

Figure 5. Decoder block diagram of harmonic magnitude VQ.

$$SD=\sqrt{\frac{1}{i_2-i_1}\sum_{n=i_1}^{i_2-1}\left[10\log_{10}\left(\frac{|m[n]|^2}{|m_q[n]|^2}\right)\right]^2} \tag{10}$$

where, i_1, i_2 are a down and up frequency value of the harmonic compnent. $m[n]$ is original spectrum. $m_q[n]$ is the spectrum synthesized appying the introduced VQ. The results are described in Figure 4. The two stages VQ showed the enhancement of 0.5～0.7 dB compared to the first method and 0.2～0.3 dB compare to the second and third method.

We used the two stages VQ for each magnitude vectors after separating low band and high band; it is considering the perceptual importance for low and high band. Figure 5 represents the block diagram of two stages VQ.

V. Transform Coding for Music

The music is synthesized by exciting an all-pole filter with an order of 10; Linear prediction (LP) was known the very efficient time-domain analysis method for low frequency[4]. The target signal for the MLT excitation analysis is the residual signal given by inverse LP filtering of input music signal. The transform coding is utilized to efficiently describe the innovative excitation for the target signal. This is accomplished by the direct encoding, in the transform domain, of the target signal from which the innovative excitation can be easily extracted. This approach preserves the principle of error minimization in the weighted-speech domain and circumvents the high complexity of analysis by synthesis approaches. The general MLT to represent the LP excitation signals is known into[12]

$$M(m)=\sum_{n=0}^{N-1}\sqrt{\frac{2}{N}}\cos(\frac{\pi}{N}(n+0.5)(m+0.5))v(n) \tag{11}$$

where

$$v(n)=w(N/2-1-n)x(N/2-1-n)+w(N/2+n)x(N/2+n)$$
$$v(n+160)=w(N-1-n)x(N+n)+w(N/2+n)x(2N-n)$$
$$w(n)=\sin\left(\frac{\pi}{2N}(n+0.5)\right)$$
$$\text{for } 0<n<N-1$$

and, IMLT and overlapp/add operation to synthesis the LP excitation signals is given by

$$M^{-1}(n)=\sum_{m=0}^{N-1}\sqrt{\frac{2}{N}}\cos(\frac{\pi}{N}(m+0.5)(n+0.5))M(m) \tag{12}$$

where

$$y(n)=w(n)M^{-1}(N/2-1-n)+w(N-1-n)M^{-1}_{pld}(n)$$
$$y(n+N/2)=w(N/2+n)M(n)-w(N/2-n)M^{-1}_{old}(N/2-1-n)$$
$$M^{-1}_{old}(n)=M^{-1}(n+160)$$
$$\text{for } 0<n<N-1$$

The MLT/IMLT can be calculated using a fast version such as[12]. The encoding process is shown in Figure 6. The peak values of MLT coefficients are extracted by select more high values than both left and right coefficient values. The peak picking process is performed considering spectral envelope effect of LP frequency response. The target coefficient is represented by

$$P(m)=|H(m)|M(m) \tag{13}$$

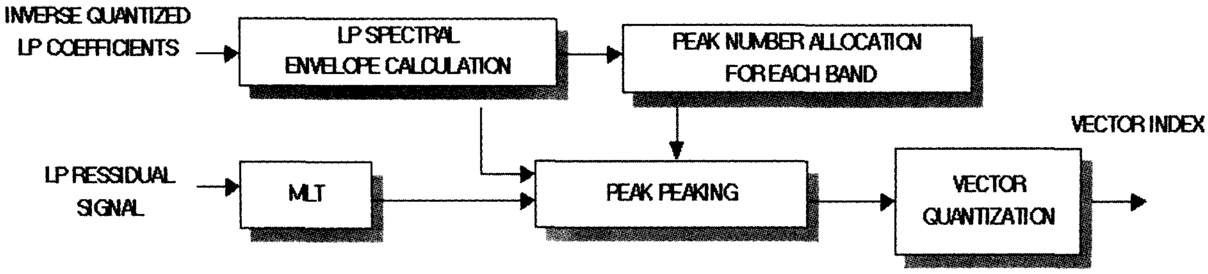

Figure 6. Encoder block diagram for music signals.

where, *P*(*m*) are the target coefficient. *H*(*m*) is the LP frequency response of input music signal. *M*(*m*) is the MLT coefficient of input music signal.

The absolute peak values are substituted from predictive values before a selection process. The differential values are quantized to satisfy a criterion minimizing spectral error with LP frequency response weight, similar to target coefficient of peak peaking process. We allocated 16 bits in the LSP parameters in music mode. The allocated bits number in music mode is small than the bits in speech mode; it is a proper number since the spectral envelope of music varies slowly than one in speech. 62 bits for the excitation parameter is assigned; MLT absolute values (20 bits), MLT sign (10 bits), MLT position (30 bits).

VI. Simulation Results

The proposed coder is implemented at 4 kbps with 20 ms frame size and 6 ms look-ahead. The bit assignment is shown in Table 1. The informal mean opinion score (MOS) including the proposed 4 kbps coder and 8 kbps QCELP were performed to evaluate the performance of proposed coder. The test was conducted with 10 participants listening to 16 sentences spoken by male and female speakers (8 male sentences, 8 female sentences). From the MOS test, we found that the 4 kbps proposed coder had better quality than 8 kbps QCELP and the enhancement of MOS 0.23 compared to the 4 kbps coder without the proposed method. Especially, the proposed

Table 1. Bit assignment of the proposed 4 kbps MTHSX coder.

Parameter	Voiced	Unvoiced	Music
LSP	24		16
V/UV/M	2		
Pitch	7	0	0
Spectral Magnitudes	5 (gain), 36 (shape)	0	0
Noise LSPs and Gains	6	0	0
Time Domain Shape	0	54	0
MLT magnitudes	0	0	20
MLT sign	0	0	10
MTL position	0	0	30
Total	80/20 ms		

Table 2. MOS test result of the proposed 4 kbps coder.

Classification	Girl	Man	Total
Original speech	4.44	4.45	4.45
8 kbps CS-ACELP	3.85	4.07	3.96
8 kbps QCELP	3.41	3.77	3.59
4 kbps Speech coding without the proposed method	3.39	3.61	3.5
4 kbps Speech coding with the proposed method	3.63	3.82	3.73

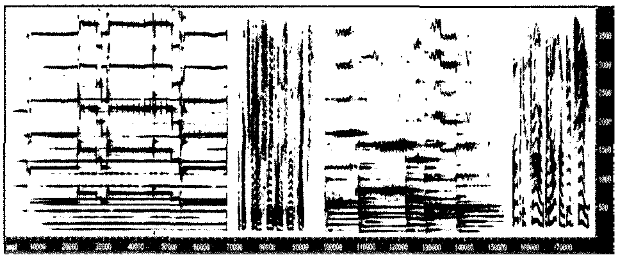

Figure 7. Original speech and music signals.

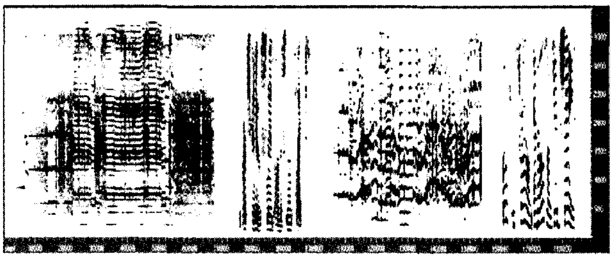

Figure 8. Reconstructed speech and music signals without the proposed method.

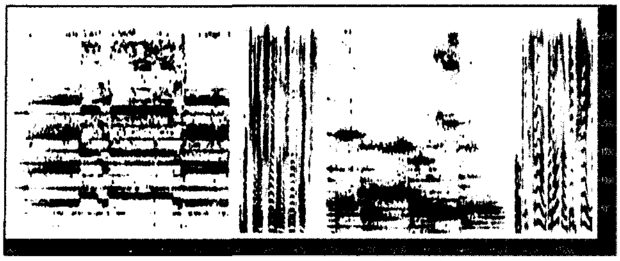

Figure 9. Reconstructed speech and music signals with the proposed method.

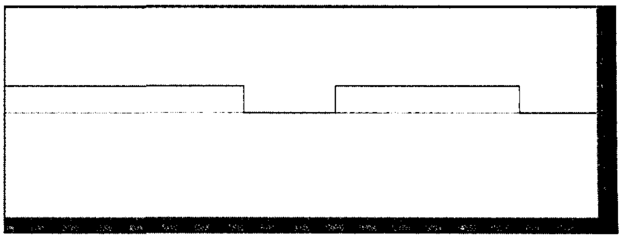

Figure 10. Speech/Music classification (speech - 0, music - 1).

coder showed an advantage for female signal. This means that is more efficient for the female speaker with large harmonic interval and many high frequency noise components. The synthesis results for the mixture signal (speech and music) are shown in Figure 7 ~ 10. The test signals include bagpipe, male, opera, female sound. We can observe the spectrum difference due to the transform coding and the harmonic coding in music signal as shown in Figure 8, 9. As shown in Figure 8, the spectral distortion introduced by harmonic coding is caused by an insufficient model for signals not to depend on a fundamental frequency like to the individual line spectrum. Only the spectral lines with large energy in the music signal were represented due to the lack of bits to be assigned.

VII. Conclusion

In this paper, the methods to complement disadvantages of harmonic coding for speech and music signal were proposed. The new cepstrum-LPC method was used to model unvoiced components for the harmonic excitation coding and the MNSTVQ method was used to quantize harmonic magnitudes that have variable dimension. A frequency domain approach based MLT excitation and adaptive peak picking process were used for efficient quantization and encoding of music excitation signal. The coder uses a combination of time domain (linear prediction) techniques to achieve the best reproduction of the original signal in a perceptual sense. The 4 kbps proposed coder showed better speech quality than that of the 8 kbps QCELP coder.

Acknowledgement

This work was supported by GRANT No. R01-2000-000-00249-0 from the Basic Research program of KOSEF (Korea Science and Engineering Foundation).

References

1. R. V. Cox, "Speech coding standards," *Speech Coding and Synthesis*, 2, W. B. Kleijn, and K. K. Paliwell Eds., Elsevier, 1995.
2. A. M. Kondoz, "Coding strategies and standards," *Digital Speech*, 5, John Wiley, 1994.
3. R. Y. Qiao, "Mixed wideband speech and music coding using a speech/music discriminator," *IEEE TENCON*, 605-608, 1997.
4. R. Lefebvre, R. Salami, C. Laflamme, and J. P. Adoul, "High quality coding of wideband audio signals using Transform Coded Excitation (TCX)," *Proc, ICASSP-94*, 1, 193-196, 1994.
5. T. Moriya, N. Iwakami, A. Jin, K. Ikeda, and S. Miki, "A design of transform coder for both speech and audio signals at 1 bit/samples," *Proc. IEEE Int. Conf. Acount., Speech, Signal Processing*, 1371-1374, 1997.
6. S. A. Ramprashad, "A two stage hybrid embedded speech/ audio coding structure," *Proc. IEEE Int. Conf. Acount., Speech, Signal Processing*, 337-340, 1998.
7. ISO/IEC JTC1/SC29/WG11, "Information technology-coding of audiovisual objects part 3: audio subpart2: parametric coding," N1903PAR, 1997
8. B. Yegnanarayana, Christophe d'Alessandro and Vassilis Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Transaction on speech and audio processing*, 6 (1), 1-11, 1998.
9. R. J. McAulay, and T. F. Quartieri, "Sinusoidal coding," *Speech Coding and Synthesis*, 4, W. B. Kleijn, and K. K. Paliwell Eds., Elsevier, 1995.
10. P. Lupini, and V. Cuperman, "Nonsquare transform vector quantization," *IEEE Signal Precessing letters*, 3 (1), January 1996
11. A. V. McCree, K. Trung, E, B, George, T. P. Barnwell

and V. Viswanathan, "A 2.4 kbit/s MELP coder candidate for the new U. S. federal standard," *Proc IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1, 200-203, May 1996.

12. H. Malvar "Fast algorithms for orthogonal and biothogonal modulated lapped transforms," *Proc IEEE Symposium, Advances in Digital Filtering and Signal Processing*, 159-163, 1998.
13. P. J. A. DeJaco, W. Gardner and C. Lee, "QCELP: north american CDMA digital cellular variable rate speech coding standard," *Proc. IEEE Workshop on speech Coding for Telecommunications*, (Sainte-Adele, Quebec), 5-6, 1993.
14. D. R. Ladd, and J. Terken, "Modelling intra- and inter-speaker pitch range variation," *Proceedings of the 13th International Congress of Phonetic Sciences Stockholm (eds. Elenius, K. & Branderud, P.)*, 2, 386-389, 1995.

[Profile]

• Jongharkk Kim

Jong hark Kim received the B. S. degree in electrical engineering and M. S degree in radio engineering from the Chungbuk National University, Korea, in 1998, 2000. He is currently a graduate student for Ph. D. degree, in radio engineering from the Chungbuk National University, Korea. His current interests include speech and audio coding, image data compression and digital signal processing.

• Jae-Hyun Shin

Jae-Hyun Shin received the B. S. degree and M.S degree in radio engineering from the Chungbuk National University, Korea, in 1997, 2000. He is currently a graduate student for Ph. D. degree, in radio engineering from the Chungbuk National University, Korea. His current interests include speech coding and recognition, digital signal processing.

• Insung Lee

Insung Lee received the B. S and M. S degrees in electronic engineering from Yonsei University, Korea, in 1983 and 1985, respectively, and the Ph. D degree in electrical engineering from Texas A&M Unversity, U. S. A, in 1992. From 1986 to 1987, he was a research engineer at the Korea Telecom Research Center, Korea. From 1989 to 1992, he was a graduate research assistant in the Department of Electrical Engineering, Texas A&M University, U. S. A. From 1993 to 1995, he was with the Signal Processing Section of the Mobile Communication Division at the Electronics and Telecommunications Research Institute (ETRI), Korea, as a senior member of technical staff. Since 1995, he has been with the Department of Radio Engineering, Chungbuk National University, as an Associate Professor. His current research interests include speech and image compression, mobile communication, adaptive filters.