LOW BIT RATE CODING FOR SPEECH AND AUDIO USING MEL LINEAR PREDICTIVE CODING (MLPC) ANALYSIS

Yoshihisa Nakatoh[†‡], *Takeshi Norimatsu*[†], *Ah Heng Low*[‡], *Hiroshi Matsumoto*[‡]

†Multimedia Development Center, Matsushita Electric Industrial Co., Ltd.
1006 Kadoma, Kadoma-shi, Osaka, 571-8501 JAPAN
‡Dept. of Electrical & Electronic Eng., Faculty of Engineering, Shinshu University.
500 Wakasato, Nagano-shi, Nagano, 380-0922, JAPAN
E-mail:nakatoh@arl.drl.mei.co.jp and matsu@sp.shinshu-u.ac.jp

ABSTRACT

This paper proposes a low bit rate coding method for speech and audio using a new analysis method named MLPC (Mel-LPC analysis). In the MLPC analysis method a spectrum envelope is estimated on a mel- or bark-frequency scale, so as to improve the spectral resolution in the low frequency band. This analysis is accomplished with about two-fold increase in computation over the standard LPC analysis.

Our coding algorithm using the MLPC analysis consists of five key parts: time frequency transformation, inverse filtering by the MLPC spectrum envelope, power normalization, perceptual weighting estimation, and the multi-stage vector quantization. In subjective experiments, we have investigated the performance of MLPC analysis method, through the evaluation of paired comparison tests between the MLPC analysis and the standard LPC one in inverse filtering. In all bit rates, almost all the listeners feel decoding signals by the MLPC analysis method is superior to the LPC one. Especially in low bit rate, there is a great difference between them.

1. INTRODUCTION

In the last few years, a significant reduction in bit rate has been demanded rapidly for wideband digital audio signal transmission and storage. This paper describes a low bit rate coding system for speech and audio using a new analysis method named MLPC (Mel-LPC analysis). Usually, speech signal production is modeled by autoregressive process. In the speech coding such as CELP[1] or the audio coding such as TwinVQ[2], LPC analysis is used to flatten the spectrum of input signal. However, in the low frequency range the spectral resolution of LPC analysis is insufficient. Because in the LPC analysis method the spectral resolution is equal at all the frequency band. As many parameters are required to represent spectrum envelope well, the bit rate can not be reduced. A linear prediction analysis on a mel- or bark-frequency scale proposed by Strube[3] is expected to be effective in speech and audio coding because of their auditory likes frequency resolution. But Strube's method needs high computational cost. So we propose MLPC analysis method[4], and apply it to speech and audio coding. In the MLPC analysis a spectrum envelope is estimated on the mel- or bark-frequency scale as in Strube's method. Our method is computationally simple (about two-fold increase) and its stability of system is guaranteed. Its performance is better than the standard LPC analysis method.

In audio coding, recently, an audio compression algorithm was proposed for MPEG-4/Audio, named "TwinVQ" [2]. In this algorithm, the MDCT coefficients of the input audio signal are divided by the LPC spectral envelope, and the flattened MDCT coefficients are encoded using interleave vector quantization. This paper describes new coding system using the MLPC analysis. In our coding system, the MLPC spectral envelope is used for flattening the MDCT coefficients and the block selective interleaved multi-stage vector quantization is used for encoding the flattened MDCT coefficients. In subjective experiments, we compared the performance of the MLPC and conventional LPC analysis, through paired comparison tests.

2. MEL LINEAR PREDICTIVE CODING (MLPC) ANALYSIS

The basic idea of all pole modeling on the warped frequency scale was proposed by Strube[3]. Strube's method is expected to be effective in speech and audio coding because of their auditory likes frequency resolution. However, this method has been rarely used in coding applications due to relatively high computational load compared to the standard LPC analysis. This paper proposes a simple and efficient time-domain technique (Mel-LPC analysis) to estimate warped predictors from input speech directly. In this study, we use a same warped inverse filter on the linear frequency axis without any prefiltering unlike Strube's method [3],

$$A_w(z) = \tilde{A}_w(\tilde{z}) = \sum_{n=0}^{p} \tilde{a}_{w,n} \tilde{z}^{-n} \tag{1}$$

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha \cdot z^{-1}} \tag{2}$$

where $\tilde{z}^{-1}$ is the first order all-pass filter. For a windowed input signal segment $x[0],\ldots,x[N-1]$, the error power is given by equation (3), and the warped predictors $\{\tilde{a}_{w,k}\}$ are estimated so as to minimize the error power over infinite time interval unlike the "covariance method" in [3].

$$\tilde{\sigma}_w^{\,2} = \sum_{n=0}^{\infty} \left\{ \sum_{k=0}^{p} \tilde{a}_{w,k} \cdot y_k[n] \right\}^2 \tag{3}$$

It should be noted that the estimated predictors $\{\tilde{a}_{w,k}\}$ are different from the predictors $\{\tilde{a}_k\}$ defined in Strube's method [3]. The warped predictors are obtained by solving for the following normal equation:

$$\sum_{j=1}^{p}\phi(i,j)\tilde{a}_{w,j} = -\phi(i,0) \qquad (i=1,\ldots,p) \tag{4}$$

where the coefficient $\phi(i,j)$ is given by

$$\phi(i,j) = \sum_{n=0}^{\infty} y_i[n]y_j[n] \tag{5}$$

using the output sequence $y_k[n]$ of the kth order all-pass filter excited by $y_0[n] = x[n]$. In terms of Parceval's theorem, $\phi(i,j)$ is proved to be equal to the autocorrelation function $\tilde{r}_w[i-j]$ of which Fourier transform is equal to the warped and frequency-weighted power spectrum, $\left|\tilde{X}\left(e^{j\tilde{\lambda}}\right)\cdot\tilde{W}\left(e^{j\tilde{\lambda}}\right)\right|^2$ as

$$\begin{aligned}\phi(i,j) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\tilde{X}\left(e^{j\tilde{\lambda}}\right)\cdot\tilde{W}\left(e^{j\tilde{\lambda}}\right)\right|^2\cos(i-j)\tilde{\lambda}\,d\tilde{\lambda}\\ &= \tilde{r}_w[i-j]\end{aligned} \tag{6}$$

where the weighting function $\left|\tilde{W}\left(e^{j\tilde{\lambda}}\right)\right|^2$ is a frequency derivative of the phase transfer function of $\tilde{z}^{-1}$, and is given by

$$\tilde{W}(\tilde{z}) = \frac{\sqrt{1-\alpha^2}}{1+\alpha\cdot\tilde{z}^{-1}} \tag{7}$$

Therefore, equation (4) becomes an autocorrelation equation as in the standard LPC analysis, and the estimated spectrum $\tilde{\sigma}_w/\tilde{A}_w(\tilde{z})$ represents the envelope of $\tilde{X}\left(e^{j\tilde{\lambda}}\right)\cdot\tilde{W}\left(e^{j\tilde{\lambda}}\right)$. If necessary, the effect of the weighting function $\tilde{W}\left(e^{j\tilde{\lambda}}\right)$ on the estimated spectrum is completely compensated by filtering $\tilde{r}_w[m]$ with the second order FIR filter, $\left[\tilde{W}(\tilde{z})\cdot\tilde{W}\left(\tilde{z}^{-1}\right)\right]^{-1}$. The resultant warped autocorrelation coefficients $\{\tilde{r}[m]\}$ lead to the same warped predictors $\{\tilde{a}_k\}$[5].

Furthermore, since $\phi(i,j)$ is a function of the difference $|i-j|$, $\phi(i,j)$ becomes equal to the sum of the following finite terms without any approximation;

$$\phi(i,j) = \tilde{r}_w[i-j] = \sum_{n=0}^{N-1} x[n]\cdot y_{(i-j)}[n] \tag{8}$$

where the output sequence $y_k[n]$ is given by

$$\begin{aligned}y_k[n] &= \alpha\cdot\left(y_k[n-1] - y_{(k-1)}[n]\right) - y_{(k-1)}[n-1]\\ &\qquad (n=0,\ldots,N-1, \quad k=1,\ldots,p)\end{aligned} \tag{9}$$

Therefore, due to the cost for computing N points of $y_k[n]$ for each i, the Mel-LPC analysis is accomplished with about two-fold increase in computation over the standard LPC analysis. This computational load is much lower than those of both "autocorrelation" and "covariance" methods in [3].

3. CODING SYSTEM

The block diagram of our coding system is illustrated in Figure 1. The encoder consists of five key parts, (1) time frequency transformation, (2) inverse filtering by the MLPC spectrum envelope, (3) power normalization, (4) perceptual weighting estimation and (5) multi-stage vector quantization (VQ). First, the input signal is transformed into the MDCT coefficients, which are flattened by the MLPC spectral envelope. The flattened MDCT coefficients are normalized by their own power. These normalized MDCT coefficients are quantized using the block selective interleaved multi-stage VQ. In every VQ stage, the adaptive selection of frequency domain used for quantization is managed. The outstanding feature of our algorithm is provided by normalization by the MLPC spectral envelope and the multi-stage VQ.

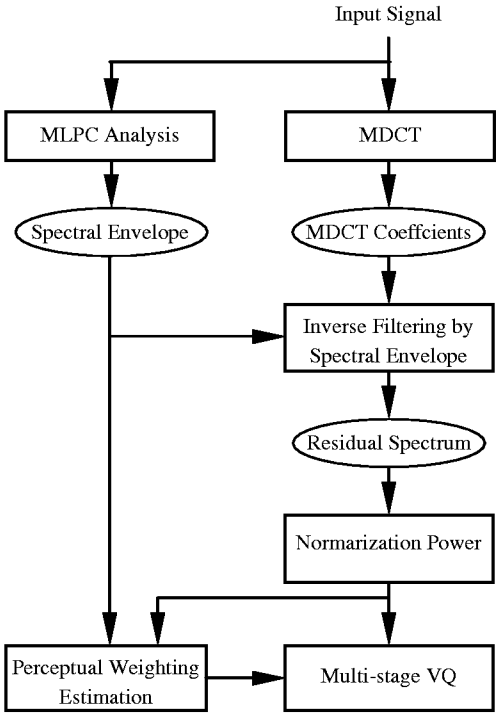

Figure 1: The block diagram of our coding system

3.1. Normalization of MDCT Coefficients

The input signal is first transformed into the frequency domain by the adaptive block size MDCT[6], and the MDCT coefficients are normalized by the spectral envelope estimated by the MLPC in SECTION 2. The MLPC spectral envelope is given by equation (10)

$$\tilde{S}\left(e^{-j\tilde{\lambda}}\right) = \frac{\tilde{\sigma}_w^{\,2}}{\left|\tilde{A}_w\left(e^{-j\tilde{\lambda}}\right)\cdot\tilde{W}\left(e^{-j\tilde{\lambda}}\right)\right|^2} \tag{10}$$

The MLPC spectral envelope in the mel- or bark-frequency domain is transformed into the linear frequency domain. The MDCT coefficients are divided by this linear spectral envelope. The MLPC coefficients are transformed into the Mel-LSP coefficients, and the quantized Mel-LSP coefficients are used in the decoder. The flattened MDCT coefficients by the MLPC spectral envelope are normalized by their own power.

3.2. Multi-Stage Vector Quantization

The flattened MDCT coefficients are progressively quantized by the block selective interleave multi-stage (3 stages) vector quantization. Figure 2 shows the block diagram of the multi-stage VQ with perceptual weighting estimation. When the window of long analysis block size (1024) is applied, the MDCT coefficients are interleaved[2], and split into sub-vectors composed of 24 elements. In case of short block size (128), the MDCT coefficients of 8 short blocks are once reassembled in order of low to high frequency. Then the reassembled vector coefficients are interleaved into the multiple sub-vectors. At first stage, each interleaved sub-vector is vector-quantized. Before entering the second stage VQ, the adaptive block selection is applied. It works to select adaptively the frequency domain that has the larger residual quantization errors and the higher perceptual sensitivity. The adaptive selection is managed according to the combination of vector quantization errors of first stage, the MLPC spectral envelope, the masking threshold curve, and a hearing threshold. The quantization errors within the selected block are interleaved and each sub-vector is vector-quantized like as a first stage VQ. The same procedure is applied in the third stage.

The adaptive block selector is designed to choose the most appropriate frequency domain that can perceptually minimize the quantization errors in the following way. First, the ideal masking threshold is calculated from MDCT coefficients utilizing the similar one to psycho-acoustic model of MPEG-1.

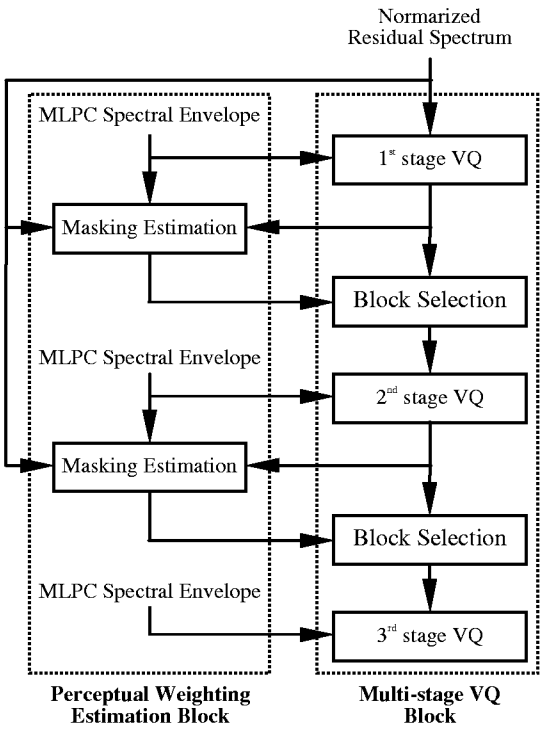

Figure 2: The block diagram of multi-stage VQ with perceptual weighting estimation

The correlation value between normalized MDCT (input to the first VQ) and the quantization errors of the first stage VQ is determined wherein the ideal masking threshold curve shows peak. Then the ideal masking threshold curve is modified by the correlation value to produce the real masking curve corresponding to a performance of quantization. The area enclosed by the quantization error curve and a masking curve is produced by every block. According to the above process, the desired block is chosen which provides the maximum area. Then the selected block becomes the input of the second VQ. The similar procedure is executed in the third stage. The MLPC spectral envelope is used as the weighting function in every stage. The every weighting function is adjusted corresponding to the distribution of the quantization error out of the previous stage VQ.

4. EXPERIMENTS

4.1. Experimental Conditions

In experiments, we used 5 kinds of the following audio samples, Pops Song, Harpsichord, Piano, Triangle and Male Speech. The specification of MLPC or LPC analysis is illustrated in Table 1. We used 4 kinds of the following bit-rates 16, 24, 32 and 40kbps. Sampling Frequency is 44.1kHz. At 16kbps the decoded sounds have of frequency bandwidth of 16kHz and at other bit rate the decoded sounds have frequency bandwidth of 20.7kHz. MDCT window length is 2048 points (long type) or 256 points (short type). The bit allocations in each bit rate are illustrated in Table 2. In the Figure, Window Type is a kind of combination of long and short MDCT window length in MDCT transformation. The bits for shift point represent the position of selected band in block selection. In each stage VQ, we used the perceptual weighted VQ method. In comparison tests we used the decoded sounds normalized by MLPC spectral envelope or LPC one in our coding system.

Table 1: Specification of the MLPC or LPC analysis

Analysis Window Length	2048points
MLPC or LPC Analysis Order p	10
Mel Scale Factor α	0.65

Table 2: Bit allocations in each bit rate

Coding Parameter	**bits/frame**(1024points)			
Mel-LSP or LSP	32			
Window Type	4			
Power	32			
Shift Point	3*2			
1st VQ(code+gain)	(4+4)*32	(6+4)*40	(8+5)*40	(8+7)*40
2nd VQ(code+gain)	(3+0)*8	(5+0)*8	(6+3)*8	(8+7)*8
3rd VQ(code+gain)	(2+0)*8	(5+0)*8	(6+3)*8	(8+7)*8
Total bits	378	554	738	914
(Bit rate)	(16kbps)	(24kbps)	(32kbps)	(40kbps)

4.2. Subjective Experiments

In subjective experiments, we investigated the performance of MLPC analysis method in the spectral normalization, through 7 level paired comparison test. First, we made a couple of decoded sounds flattened by MLPC spectral envelope or LPC one in our coding system and we presented it at random to 8 listeners (including acoustic specialists). Next, all listeners decide 7 level comparison scores after listing a pair of sounds. Test item is " Which source do you feel higher quality? ". Table 3 is preference score (average of all listeners) of the MLPC analysis method in comparison with the standard LPC analysis method in inverse filtering part. Its confidence level of 95% is 0.3 to 0.5. In all bit rates, almost all the listeners feel that the decoded sounds by the MLPC analysis method are superior to the decoded sounds by the LPC one. Especially in 16kbps, there is a great difference between them. And we have found that the effect of the proposed MLPC analysis is larger than the standard LPC method about Pops Song, Piano and Male Speech.

Table 3: Preference score of the MLPC analysis in comparison with the standard LPC analysis in inverse filtering by the spectrum envelop (average of all listeners)

Audio Sample	**bit rate**[kbps]			
	16	24	32	40
Pops Song	1.00±0.3	0.36±0.4	0.36±0.5	0.21±0.4
Harpsichord	0.29±0.3	0.14±0.4	-0.29±0.5	0.29±0.4
Piano	0.79±0.4	0.21±0.4	0.57±0.5	0.57±0.4
Triangle	0.36±0.4	0.43±0.4	-0.14±0.5	0.29±0.4
Male Speech	0.79±0.3	0.36±0.4	0.07±0.4	0.14±0.3

5. DISCUSSION

We compared the proposed MLPC spectral envelope with the standard LPC one to investigate the performance of MLPC analysis method. Fig.3 shows the spectral envelopes of male speech at analysis order of p=16. In this figure, the dotted line is the LPC spectral envelope and the solid line is the MLPC one. The FFT spectrum is illustrated in this figure, too. The horizontal axis represents mel- frequency scale to make the detail of spectrum at low frequency band clear. It is clear that the spectrum by MLPC analysis method is much better than the spectrum by LPC one. Especially, it is remarkable at low frequency band under 5kHz.

6. CONCLUSION

We proposed a low bit rate coding method for speech and audio using the MLPC analysis. The outstanding feature of our coder is provided by inverse filtering by the MLPC spectral envelope and the block selective interleave multi-stage VQ with perceptual weighting estimation. The MLPC analysis is the method to estimate a spectrum envelope on a mel- or bark-

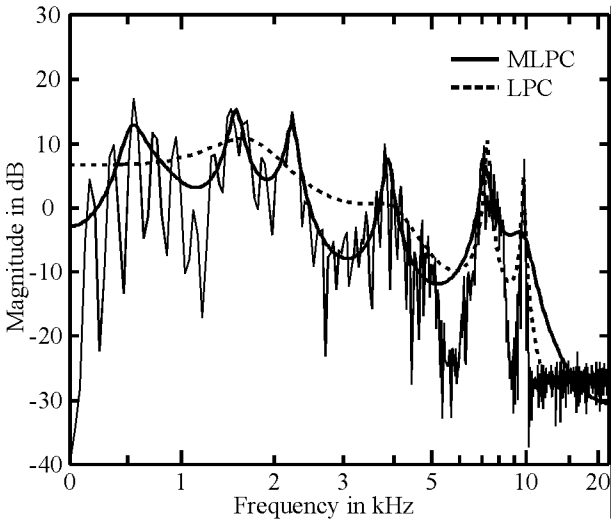

Figure 3: Comparison the MLPC spectral envelope with the standard LPC one about male speech at analysis order of p=16

frequency scale and this paper proposed a simple and efficient time-domain technique to estimate warped predictors from an input speech directly. We have investigated the performance of the MLPC analysis in inverse filtering through 7 level paired comparison tests. In all bit rates, almost all the listeners feel that decoded sounds by the MLPC analysis method is superior to the decoded sounds by the LPC one. Especially in low bit rate, there is a great difference between them. In the future, we will study improvement of our coding algorithm and apply to wideband speech coding.

REFERENCES

[1] M.R.Schroeder and B.S.Atal, "Code Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates," Proc.of ICASSP85, pp.937-940, 1985.

[2] N.Iwakami and T.Moriya, "High-quality Audio Coding at less than 64 kbit/s by Using TwinVQ," Proc.of ICASSP95, vol.5, pp.3095-3098, 1995.

[3] H.W.Strube, "Linear prediction on a warped frequency scale," J.of Acoust.Soc.America, vol.68, no.4, pp.1071-1076, 1980.

[4] Y.Nakatoh, T.Norimatsu, A.H.Low and H.Matsumoto, "An Improved Estimation Algorithm of Spectrum Envelope for Audio Coding," Spring Meeting of ASJ, 2-7-6, pp.245-146, 1998.

[5] H.Matsumoto, Y.Nakatoh and Y.Furuhata, " An Efficient Mel-LPC Analysis Method for Speech Recognition," Proc.of ICSLP98, 1998.

[6] M.Iwadare, A.Sugiyama, F.Hazu, A.Hirano and T.Nishitai, "A 128 kb/s Audio CODEC Based on Adaptive Transform Coding with Adaptive Block Size MDCT," IEEE JSAC, vol.10, no.1, pp.138-144, 1992.