HARMONIC AND NOISE CODING OF LPC RESIDUALS WITH CLASSIFIED VECTOR QUANTIZATION

Masayuki Nishiguchi and Jun Matsumoto

InfoCom Products Company, Sony Corporation
1-7-4 Konan, Minato-ku, Tokyo, Japan

ABSTRACT

An efficient coding scheme for Linear Predictive Coding (LPC) residuals is proposed based on harmonic and noise representation. New features of the scheme include classified vector quantization of the spectral envelope of LPC residuals with a weighted distortion measure. The improvement in performance obtained by classifying codebooks based on a voiced / unvoiced (V/UV) decision is shown. Sequences of the short-term rms power of time domain waveforms are also vector quantized and transmitted for unvoiced signals. A fast synthesis algorithm for voiced signals using an FFT is also presented, which reduces the high complexity of the direct sinusoidal synthesis method with interpolated magnitudes and phases. Informal listening tests indicate that, in combination with a known LSP quantization technique, this residual coding scheme provides good communication quality at a total bit rate of less than 2.0Kbps.

1. INTRODUCTION

The Code-Excited Linear Prediction (CELP) Algorithm [1] is widely used at bit rates of around 4~16Kbps and provides natural speech. When the bit rate is below 3Kbps, however, the quality deteriorates, perhaps because of the nature of the waveform matching, which requires phase reconstruction as well. On the other hand, for bit rates below 3~4Kbps, a class of sinusoidal coding that includes Harmonic coding [2], Sinusoidal Transform Coding (STC) [3], and Multi Band Excitation (MBE) [4][6] is known to provide good communication quality mainly due to the smooth reconstruction of voiced signals. To obtain natural quality, however, these coders also have to use phase information as well as spectral magnitude, which results in higher rates. A simple but effective compromise that roughly reproduces the phase without any extra bits is a scheme combining linear predictive coding (LPC) and Harmonic coding, i.e., the harmonic representation of LPC residuals. Harmonic coding is an effective and natural way of discarding the phase of residuals and thus lowering the bit-rate without causing any discontinuity.

The principle of the proposed algorithm is similar to that of Time-Frequency Interpolation (TFI)[5]. The major differences are that our method extracts the harmonic structure from the power spectrum obtained from an FFT with a fixed length and interval, without requiring any pitch synchronous waveform extraction. Also, a unified quantization strategy for both voiced and unvoiced segments is used, thereby simplifying the overall coding scheme. This is enabled by classified vector quantization of the spectral envelope, and vector quantization of a sequence of short-term rms power values of the time domain waveform.

This paper is organized as follows. Section 2 presents the basic coder algorithm including spectral and short-term rms quantization. The fast harmonic synthesis algorithm is explained in Section 3. Coder parameters and experimental results are shown in Section 4.

2. PROPOSED CODER SCHEME

Fig.1 shows the overall structure of the encoder and decoder. Speech input at a sampling rate of 8KHz is formed into frames with a length and interval of 256 and 160 samples, respectively. LPC analysis is carried out using windowed input data over one frame. LPC residual signals are computed by the inverse filtering of input data using quantized and interpolated LSP parameters. The residual signals are then fed into the pitch and spectral magnitude estimation block, where the spectral envelopes for both voiced and unvoiced signals are estimated in the same manner as in the MBE coder [4] except that only a one-bit V/UV decision is used per frame. The spectral envelope and a sequence of short-term rms power values of the residual signals over one frame are then vector quantized. Detailed configurations are described below.

2.1. Classified vector quantization of spectral envelope with a weighted distortion measure

In order to vector quantize a spectral envelope composed of a variable number of harmonics, the spectral vector is converted to a fixed-dimension vector. We use band-limited interpolation by a polyphase filter bank for this dimensional conversion [6].

A fixed-dimension (=44) spectral vector is then quantized. In order to reduce the memory requirements and search complexity while maintaining a high performance, a Multi-Stage (2-stage) Vector Quantization (MSVQ) scheme is employed for the spectral shape together with a scalar quantizer for the gain, as shown in Fig.2. The weighted distortion measure, D, below is used throughout the design and search of the shape and gain codebooks.

$$D = \| \frac{1}{\|\mathbf{x}\|} \mathbf{W}\mathbf{H}(\mathbf{x} - g(\mathbf{s}_0 + \mathbf{s}_1)) \|^2 , \qquad (1)$$

0-7803-2431-5/95 $4.00 © 1995 IEEE

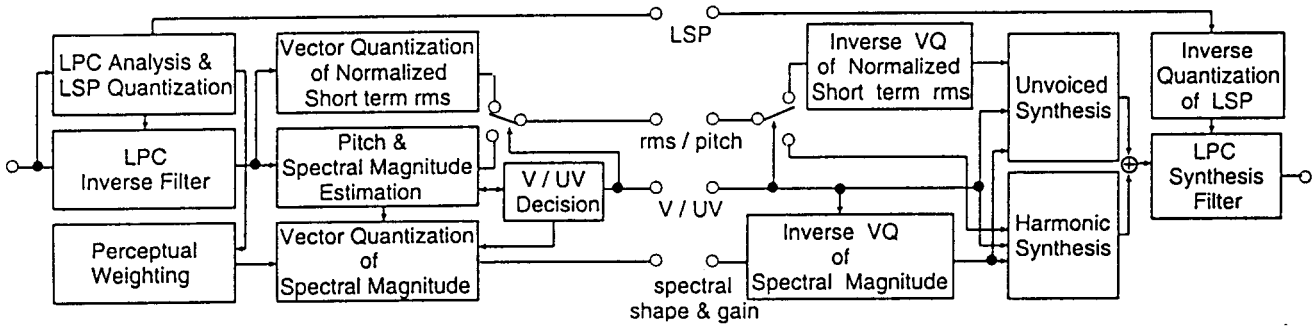

Figure 1. Overall structure of the coder and decoder

where $\mathbf{x}$ is a source vector, $\mathbf{s_0}$ is the output of Shape Codebook-0 (CB-0), $\mathbf{s_1}$ is the output of Shape Codebook-1 (CB-1), and g is the output of the gain codebook. The diagonal components of the matrices $\mathbf{H}$ and $\mathbf{W}$ are the magnitudes of the frequency response of the LPC synthesis filter and the perceptual weighting filter, respectively. An optimal MSVQ design [7][8] based on the Generalized Lloyd Algorithm (GLA)[9] is used. Summing up the distortion from all the frames which use the j-th shape vector $\mathbf{s}_{kj}$ in Codebook-k (k=0,1), and taking the derivative of the summed distortion with respect to $\mathbf{s}_{kj}$ and setting the result equal to zero, we obtain the new centroid, $\mathbf{u}_{kj}$, for $\mathbf{s}_{kj}$:

$$\mathbf{u}_{kj} = (\sum_{i \in j} g_i{}^2 \mathbf{A}_i{}^T \mathbf{A}_i)^{-1} \sum_{i \in j} g_i \mathbf{A}_i{}^T \mathbf{A}_i (\mathbf{x}_i - g_i \mathbf{s}_{1-k\,i}) , \quad (2)$$

where

$$\mathbf{A}_i = \mathbf{W}_i \mathbf{H}_i / \|\mathbf{x}_i\| . \quad (3)$$

The suffix i denotes the frame number that selects $\mathbf{s}_{kj}$ from Codebook-k. $\mathbf{x}_i$; $\mathbf{W}_i$, and $\mathbf{H}_i$ are the source vector and associated weights; and g_i and $\mathbf{s}_{1-k\,i}$ are the codewords selected in frame i. Similarly, the new centroid v_j for the j-th gain codeword is given by:

$$v_j = \sum_{i \in j} \mathbf{x}_i{}^T \mathbf{A}_i{}^T \mathbf{A}_i (\mathbf{s}_{0i} + \mathbf{s}_{1i}) / \sum_{i \in j} \|\mathbf{A}_i (\mathbf{s}_{0i} + \mathbf{s}_{1i})\|^2 , \quad (4)$$

where the suffix i denotes the frame number that selects the j-th gain codeword.

Fig.3 shows the average of the source vector $\mathbf{x}$, and the average of the weight $\mathbf{WH}$ for voiced and unvoiced segments taken from about 6 minutes of training data. As can be seen, there is not much difference in the average shape of the source for V and UV. However, the average weight strongly depends on the V/UV decision. This means that it would be better to use different classes of bit allocations over a spectral vector for V and UV. We therefore introduced classified VQ [10], in which all three codebooks (CB-0, CB-1, gain) are switched based on the V/UV decision. Note that the V/UV flag must be transmitted in any case. Thus, the sizes of the codebooks for V and UV remain the same as those for a fixed codebook scheme and no extra bits are consumed.

Using a set of typical speech sequences outside the training data, the segmental SNR between synthesized speech segments with and without spectral quantization was measured. Compared with a fixed codebook scheme, the average segmental SNR of the classified codebook scheme was 1.3 dB higher (15.7dB → 17.0dB) for a codebook size of 5 bits each for CB-0, CB-1, and gain.

2.2. Gain adjustment of unvoiced residual signals

When a segment is unvoiced, no pitch information is used. Instead, the coder sends information describing the envelope of the LPC residual waveform. The short-term rms power is computed over every 4-ms period, and an 8-dimensional vector composed of a sequence of eight rms normalized by a frame rms is vector quantized in each frame. In the decoder, the IFFT of the noise colored by the quantized spectral vector $\hat{\mathbf{x}}$ is computed first. The time domain waveform is then gain adjusted so that the distribution of its short-term rms power over a frame equals that of the transmitted one. Fig.4 shows the operation of this unvoiced synthesis.

Together with the classified quantization of the spectral envelope described above, this rms adjustment provides a very simple but effective representation of unvoiced speech, making the complex structure of the CELP-type approach unnecessary.

It should be noted that this gain adjustment improves the quality of unvoiced signals not only in the LPC residual coding scheme used here, but also in a regular MBE coder.

3. HARMONIC SYNTHESIS BY IFFT

One drawback of Harmonic coding is the high complexity of the synthesizer. If we compute the voiced output, $v(n)$, directly from the equation

$$v(n) = \sum_m A_m(n) cos(\theta_m(n)) \quad (5)$$

$$(0 \le m < M,\ 0 \le n < N) ,$$

with interpolated magnitudes $A_m(n)$ and phases $\theta_m(n)$ [4], then the complexity is on the order of γNM, where n is a

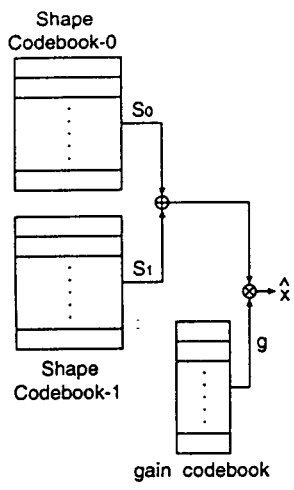

Figure 2. Two-stage VQ scheme

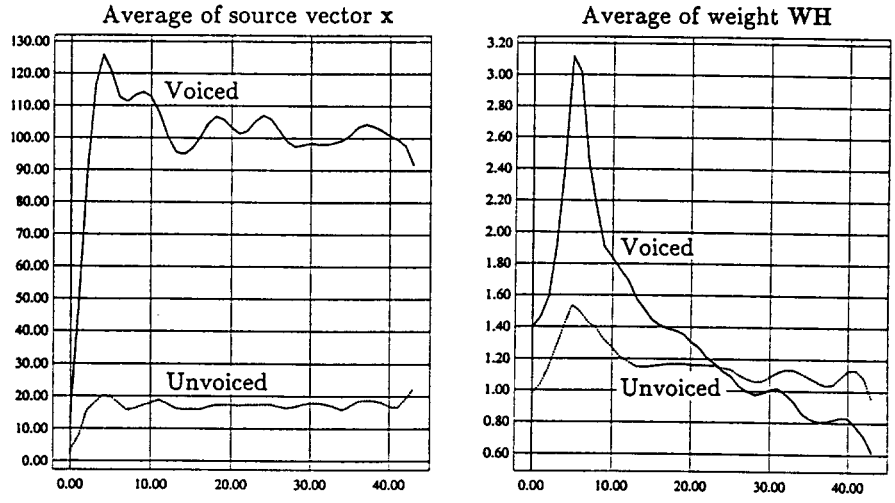

Figure 3. Averages of x and WH

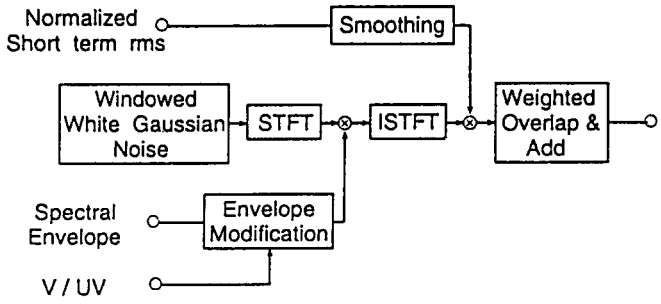

Figure 4. Synthesis of unvoiced residual signals

discrete time index and m is a harmonic index. γ is a constant related to the interpolation of magnitude and phase. Typically, $N = 160$, $M = 64$, and $\gamma = 5$. In order to reduce the complexity, we developed a fast synthesis method using an IFFT and sampling rate conversion, in which $A_m(n)$ and $\theta_m(n)$ are linearly interpolated.

Suppose at the k^{th} frame we have a spectrum with M_1 harmonics, with the magnitude of each being A_m $(0 \leq m < M_1)$. The pitch lag expressed in terms of the number of samples is $2M_1$. Appending zeros to this array of A_m yields a new array with 2^b components ranging from 0 to π. The number b can be arbitrarily chosen so that $M \leq 2^b$. The same processing is done on the array of phase data. The phase data used here are generated from those of the previous frame assuming that the fundamental frequency is linearly interpolated. This is described as phase prediction in [4][2]. 2^{b+1}-point IFFT is applied to these arrays of magnitude and phase data with the constraint that the results be real numbers. Now we have an over-sampled version of the time domain waveform over a one-pitch period. Let this be $w_1(i)$ $(0 \leq i < 2^{b+1})$. 2^{b+1} points are used to express the one-pitch period of the waveform, whereas the actual pitch is $2M_1$ since the over-sampling ratio ov_1 is

$$ov_1 = 2^b / M_1 \,. \tag{6}$$

Similarly, we can get another one-pitch period of the waveform at the $k+1^{th}$ frame, which has an over-sampling ratio of

$$ov_2 = 2^b / M_2 \,, \tag{7}$$

where the pitch lag is $2M_2$. Let this waveform be $w_2(i)$ $(0 \leq i < 2^{b+1})$. Here we define the function $f(n)$, which maps the time index n from the original sampling version to the over-sampled version under the condition that the pitch period is linearly interpolated, to be

$$f(n) = \int_0^n (ov_1 \frac{N-t}{N} + ov_2 \frac{t}{N}) dt \,. \tag{8}$$

The number of over-sampled data needed to reconstruct a waveform of length N at the original sampling rate is at most L:

$$L = nint(f(N)) = nint(\frac{N}{2}(ov_1 + ov_2)) \,, \tag{9}$$

where $nint(x)$ returns the nearest integer of x. Cyclicly extending $w_1(i)$ and $w_2(i)$, we obtain waveforms $\tilde{w}_1(l)$ and $\tilde{w}_2(l)$ of length L:

$$\tilde{w}_1(l) = w_1(mod(l, 2^{b+1})) \qquad (0 \leq l < L) \,, \tag{10}$$

$$\tilde{w}_2(l) = w_2(mod(offset + l, 2^{b+1})) \quad (0 \leq l < L) \,, \tag{11}$$

where

$$offset = 2^{b+1} - mod(L.\, 2^{b+1}) \,, \tag{12}$$

and $mod(x, y)$ returns the remainder of x divided by y. These two waveforms, $\tilde{w}_1(l)$ and $\tilde{w}_2(l)$, from the spectra of the k^{th} and $k+1^{th}$ frames have the same "pseudo" pitch $(= 2^{b+1})$ and they are aligned. So, simply adding these two

waveforms using appropriate weights produces the result $w(l)$:

$$w(l) = \frac{L-l}{L}\tilde{w}_1(l) + \frac{l}{L}\tilde{w}_2(l) \qquad (0 \le l < L)\,, \quad (13)$$

where each A_m is linearly interpolated between the adjacent frames. Lastly, $w(l)$ has to be re-sampled so that the resulting waveform can be expressed at the original uniform sampling rate. This operation brings the waveform back from the "pseudo" pitch domain to the real pitch domain as well. In principle, the re-sampling operation is just

$$v'(n) = w(f(n)) \qquad (0 \le n < N)\,. \quad (14)$$

Usually $f(n)$ does not return integer values. So, $v'(n)$ is obtained by linearly interpolating $w(\lceil f(n) \rceil)$ and $w(\lfloor f(n) \rfloor)$. For a more general formulation, a higher order interpolation could be used. $\lceil x \rceil$ and $\lfloor x \rfloor$ denote the smallest integer greater than or equal to x, and the largest integer less than or equal to x, respectively.

$v'(n)$ is a very good approximation of $v(n)$ in Eq.(5) and reduces the complexity to the order of $\alpha 2^b(b+1)+\beta N$. α and β are constants related to the IFFT and linear interpolation, respectively. Listening experiments demonstrated a b of 6 to be sufficient, and synthesized speech was indistinguishable from that obtained by the direct method of Eq.(5). The average of the segmental SNR between the two results was about 32 dB. Typically, $\alpha = 7$ and $\beta = 12$, and thus the complexity is less than 1/10 that of the direct method.

4. SIMULATION AND PERFORMANCE

Table 1 shows the bit allocations of the coder. In combination with the LSP matrix quantization technique used in the PDC half-rate standard [11], this residual coding scheme provides good communication quality at a total bit rate of 1.975 Kbps.

The performance of the proposed 1.975-Kbps coder under no background noise conditions was compared with 6.7-Kbps VSELP (PDC full-rate standard [11]) in an informal listening test. 16 Japanese sentences spoken by several male and female speakers were compared by 30 listeners. Table 2 shows the results. Though this is only a limited aspect of the coder performance, it indicates that the potential performance is comparable to 6.7-Kbps VSELP.

5. CONCLUSIONS

In summary, it has been proven that a classified VQ scheme for the spectral envelope improves the quantization performance of LPC residuals in harmonic/noise representation. This enables a common quantization strategy to be applied to both voiced and unvoiced signals in combination with a gain adjustment technique.

Also a simple but effective fast algorithm for harmonic synthesis was developed, and the necessary constant values were given.

Parameters	Bit Allocations
LSP	31bit/40ms
Pitch(Voiced)/	
Short-term rms(Unvoiced)	8bits/20ms
V/UV	1bit/20ms
Spectral shape	10bits/20ms
Gain	5bits/20ms
Total	79bits/40ms

Table 1. Bit Allocations for a 1.975 Kbps coder

Prefer Proposed Coder	No Preference	Prefer VSELP
38%	26%	36%

Table 2. Comparison of a 1.975 Kbps coder with 6.7 Kbps VSELP

REFERENCES

[1] M.R.Schroeder and B.S.Atal, "Code-Excited Linear Predictive (CELP): High-Quality Speech at Very Low Bit Rates," Proc.ICASSP-85,pp.937-940,1985

[2] J.S. Marques, L.B. Almeida, and J.M. Tribolet, "Harmonic coding at 4.8 kb/s," Proc. ICASSP-90,pp.I-17-20,1990

[3] R.J.McAulay and T.F.Quatieri, "Speech Analysis /Synthesis Based on a Sinusoidal Representation," IEEE Trans. ASSP, Vol.34, No.4, pp.744-754,Aug. 1986

[4] D.W.Griffin and J.S.Lim, "Multiband Excitation Vocoder," IEEE Trans. ASSP, Vol.36, pp.1223-1235, Aug. 1988

[5] Y.Shoham, "High-Quality Speech Coding at 2.4 to 4.0 Kbps based on Time-Frequency Interpolation," Proc. ICASSP-93,pp.II-167-170,Apr.1993

[6] M.Nishiguchi, J.Matsumoto, S.Ono, R.Wakatsuki, "Vector Quantized MBE with Simplified V/UV Division at 3.0Kbps," Proc. ICASSP-93, pp.II-151-154, Apr.1993

[7] T.Moriya, "Two-Channel Conjugate Vector Quantizer for Noisy Channel Speech Coding," IEEE JSAC, Vol.10, pp.866-874,June 1992

[8] W.-Y.Chan, *Product Code Vector Quantization Methods with Application to High Fidelity Audio Coding,* PhD thesis, University of California Santa Barbara, 1991

[9] Y.Linde, A.Buzo, and R.M.Gray, "An algorithm for vector quantizer design," IEEE Trans. Comm., vol. COM-28, pp.84-95, Jan. 1980

[10] A.Gersho, R.M.Gray, *Vector Quantization and Signal Compression,* Kluwer Academic Publishers, 1992

[11] Research and Development Center for Radio Systems, *Personal Digital Cellular Telecommunication System RCR Standard (RCR STD27C),* Japan, 1994