DISPERSION PHASE VECTOR QUANTIZATION FOR ENHANCEMENT OF WAVEFORM INTERPOLATIVE CODER

Oded Gottesman

Signal Compression Laboratory
Department of Electrical and Computer Engineering
University of California
Santa Barbara, California 93106, USA
E-mail: oded@kane.ece.ucsb.edu

ABSTRACT

This paper presents an efficient *analysis-by-synthesis* vector quantizer for the dispersion phase of the excitation signal which was used to enhance a *waveform-interpolative* coder. The scheme can be used to enhance other harmonic coders, such as the *sinusoidal-transform coder* and the *multiband-excitation* coder. The scheme incorporates perceptual weighting, and does not require any phase unwarping. The proposed quantizer achieves a segmental signal-to-noise ratio of up to 14dB for as low as 6-bit quantization. Subjective testing shows improvement in synthesized speech quality using the quantized phase over a male speaker extracted phase. The improvement was larger for female speakers.

1. INTRODUCTION

Recently, there has been growing interest in developing toll-quality speech coders at rates of 4kbps and below. The speech quality produced by waveform coders such as *code-excited linear predictive* (CELP) coder [[1]] degrades rapidly at rates below 5kbps. On the other hand, parametric coders such as the *Waveform-interpolative* (WI) coder [[5]-[15]], the *sinusoidal-transform coder* (STC) [[2],[3]], and the *multiband-excitation* (MBE) coder [[4]] produce good quality at low rates, but they do not achieve toll quality. In parametric coders the phase information is commonly not transmitted, and this is for two reasons: first, the phase is of secondary perceptual significance; and second, no efficient phase quantization scheme is known. WI coders [5-[15]] typically use a fixed phase vector for the *slowly evolving waveform* (SEW), for example, in [[10],[15]] fixed male speaker extracted phase was used. On the other hand, Waveform coders such as CELP [[1]], by directly quantizing the waveform, implicitly allocate an excessive number of bits to the phase information - more than is perceptually required.

In the past, phase modeling and quantization was investigated. In [[16]] a random phase codebook was used at a relatively high number of phase quantization bits. In [[17],[18]] a non-causal all-pole filter's phase model was discussed, but quantization was not optimized. Such a model is occasionally limited in matching the physiological excitation's phase. In addition, none of the above methods have incorporated perceptual weighting.

In this work, we propose a novel, efficient *analysis-by-synthesis* (AbS) quantization scheme for the phase at a very low bitrate, which can be used for parametric coders as well as for waveform coders. The proposed quantizer has been implemented as part of a WI system to quantize the SEW phase, and its performance has been investigated.

This paper is organized as follows. In Section 2 we discuss the dispersion phase quantizer, with emphasis on the distortion measure, and the respective optimal codebook design. In Section 3 we then describe the objective results obtained by the designed codebook, as well as results of a subjective test which compares the proposed quantizer, using only 4 bits, with a fixed phase vector extracted from a male speaker. Finally, we summarize our work.

2. PHASE QUANTIZATION

The dispersion-phase quantization scheme is illustrated in Figure 1. Consider a pitch cycle which is extracted from the residual signal, and is cyclically shifted such that its pulse is located at position zero. Let its DFT be denoted by $\mathbf{R}$; the resulting DFT phase is the *dispersion phase*, φ, which determines, along with the magnitude $|\mathbf{R}|$, the waveform's pulse shape. After quantization, the components of the quantized magnitude vector, $|\hat{\mathbf{R}}|$, are multiplied by the exponential of the quantized phases, $\hat{\varphi}(k)$, to yield the quantized waveform DFT, $\hat{\mathbf{R}}$, which is subtracted from the input DFT to produce the error DFT. The error DFT is then transformed to the perceptual

This work was supported in part by the University of California MICRO program, ACT Networks, Inc, Advanced Computer Communications, Cisco Systems, Inc., DSP Group, Inc., Fujitsu Laboratories of America, Inc., Hughes Electronics Corp., Intel Corp., Lucent Technologies, Inc., Nokia Mobile Phones, Qualcomm, Inc., Rockwell International Corp., Sun Microsystems, Inc., and Texas Instruments, Inc.

domain by weighting it by the combined synthesis and weighting filter W(z). The encoder searches for the phase that minimizes the energy of the perceptual domain error, allowing a refining cyclic shift of the input waveform during the search, to eliminate any residual phase shift between the input waveform to the quantized waveform.

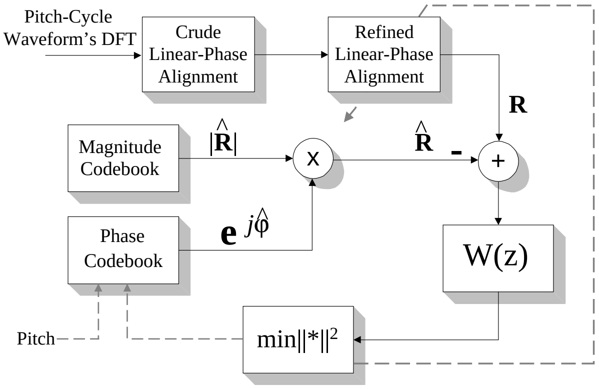
Pitch-Cycle Waveform's DFT
Crude Linear-Phase Alignment
Refined Linear-Phase Alignment
R
Magnitude Codebook
$|\hat{\mathbf{R}}|$
x
$\hat{\mathbf{R}}$
-
+
Phase Codebook
$\mathbf{e}^{j\hat{\phi}}$
W(z)
Pitch
min||*||^2

Figure 1. Block diagram of the AbS dispersion phase's vector quantization.

2.1. Waveform Matching Distortion

Phase dispersion quantization aims to improve waveform matching. Efficient AbS quantization can be obtained by using the perceptually weighted distortion measure:

$$D_w = \frac{1}{2\pi}\int_0^{2\pi} \left| r_w(\phi) - \hat{r}_w(\phi) \right|^2 d\phi$$

$$= \sum_{k=-K}^{K} W_{kk} \left| R_k - \hat{R}_k \right|^2 = \frac{1}{P}\sum_{k=0}^{P-1} W_{kk} \left| R(k) - \hat{R}(k) \right|^2 \qquad (1)$$

where

$r_w(\phi)$ - weighted (synthesized) band-limited pulse prototype,

$\hat{r}_w(\phi)$ - weighted (synthesized) quantized pulse prototype,

$R_k, \hat{R}_k$ - Fourier series coefficients of the non-weighted pulse prototypes $r(\phi)$, and $\hat{r}(\phi)$, respectively,

$R(k)$, $\hat{R}(k)$ - DFT coefficients of $r(\phi)$, and $\hat{r}(\phi)$, respectively,

P - pitch period in samples,

K - number of harmonics, $K = \left\lfloor \frac{P}{2} \right\rfloor$

W_{kk} - combined spectral-weighting and synthesis of the k-th harmonic given by:

$$W_{kk} = \left| \frac{A(z/\gamma_1)}{A(z)A(z/\gamma_2)} \right|^2_{z=e^{j(\frac{2\pi}{P})k}} \qquad (2)$$

where $A(z)$ is the LPC polynomial, and the spectral weighting parameters satisfy:

$$0 \le \gamma_2 < \gamma_1 \le 1 \qquad (3)$$

We can rewrite equation (1) in vector notation:

$$D_w(\mathbf{R}, \hat{\mathbf{R}}) = (\mathbf{R} - \hat{\mathbf{R}})^H \mathbf{W} (\mathbf{R} - \hat{\mathbf{R}}) / K \qquad (4)$$

where,

$$\mathbf{R} = \left[R(1), \ldots, R(K) \right]^T \qquad (5)$$

is the input DFT vector,

$$\hat{\mathbf{R}} = \left[\hat{R}(1), \ldots, \hat{R}(K) \right]^T \qquad (6)$$

is the quantized DFT vector, and

$$\mathbf{W} = diagonal\left\{ W_{kk} \right\} \qquad (7)$$

The magnitude is perceptually more significant than the phase; and should, therefore, be quantized first. Since phase is quantized to fewer bits, unless magnitude is quantized first, the magnitude spectral matching is unnecessarily sacrificed for excessive waveform matching. Additional justification for the above is provided from computational complexity considerations. For the above distortion measure, the quantized phase vector is given by:

$$\hat{\varphi} = \arg\min_{\hat{\varphi}_i} \left\{ D_w(\mathbf{R}, \mathbf{e}^{j\hat{\varphi}_i} |\hat{\mathbf{R}}|) \right\}$$

$$= \operatorname{argmin}_{\hat{\varphi}_i} \left\{ (\mathbf{R} - \mathbf{e}^{j\hat{\varphi}_i} |\hat{\mathbf{R}}|)^H \mathbf{W} (\mathbf{R} - \mathbf{e}^{j\hat{\varphi}_i} |\hat{\mathbf{R}}|) \right\}$$

$$= \arg\max_{\hat{\varphi}_i} \left\{ \int_0^{2\pi} r_w(\phi) \hat{r}_w(\hat{\varphi}_i, \phi) d\phi \right\} \qquad (8)$$

where i is the running phase codebook index, and the respective phase exponent matrix is given by:

$$\mathbf{e}^{j\hat{\varphi}_i} = diagonal\left\{ e^{j\hat{\varphi}_i(k)} \right\} \qquad (9)$$

Equivalently, the quantized phase vector can be simplified to:

$$\hat{\varphi} = \arg\max_{\hat{\varphi}_i} \left\{ \sum_{k=1}^{K} W_{kk} \left| R(k) \right| \left| \hat{R}(k) \right| \cos(\varphi(k) - \hat{\varphi}(k)_i) \right\} \quad (10)$$

where $\varphi(k)$ is the phase of, $R(k)$, the k-th input DFT coefficient.

The average global distortion measure for M vector set is:

$$D_{w,Global} = \frac{1}{M} \sum_{\substack{m=\{Data \\ Vectors\}}} D_w(\mathbf{R}_m, \mathbf{e}^{j\hat{\varphi}_m} \left| \hat{\mathbf{R}} \right|_m) \quad (11)$$

$$= \frac{1}{M} \sum_{\substack{m=\{Data \\ Vectors\}}} \frac{1}{K_m} \sum_{k=1}^{K_m} \mathbf{W}_{kk,m} \left| R(k)_m - e^{j\hat{\varphi}(k)_m} \left| \hat{R}(k) \right|_m \right|^2$$

2.2. Centroid Equations

The centroid equation [[19]] of the k-th harmonic's phase, for the j-th cluster, which minimizes the global distortion in equation (11), is given by:

$$\hat{\varphi}(k)_{j^{th}-cluster}$$

$$= \operatorname{atan} \left[\frac{\sum_{m=\{j^{th}-cluster\}} \frac{1}{K_m} W_{kk,m} \left| \hat{R}(k)_m \right| \left| R(k)_m \right| \sin(\varphi(k)_m)}{\sum_{m=\{j^{th}-cluster\}} \frac{1}{K_m} W_{kk,m} \left| \hat{R}(k)_m \right| \left| R(k)_m \right| \cos(\varphi(k)_m)} \right] \quad (12)$$

These centroid equations use trigonometric functions of the phase, and therefore do not require any phase unwarping.

2.3. Variable Dimension VQ

The phase vector's dimension depends on the pitch period and, therefore, a variable dimension VQ has been implemented. In our WI system the possible pitch period value was divided into eight ranges, and for each range of pitch period an optimal codebook was designed such that vectors of dimension smaller than the largest pitch period in each range are zero padded.

Pitch changes over time cause the quantizer to switch among the pitch-range codebooks. In order to achieve smooth phase variations whenever such switch occurs, overlapped training clusters were used.

3. EXPERIMENTAL RESULTS

3.1. Objective Results

Our phase-quantization scheme has been implemented as a part of WI coder, and used to quantize the SEW phase. The objective performance of the suggested phase VQ has been tested under the following conditions:

- Phase Bits: 0-6 every 20ms, a bitrate of 0-300 bit/second.
- 8 pitch ranges were selected, and training has been performed for each range.
- Modified IRS (MIRS) [[20]] filtered speech (Female+Male)
  - Training Set: 99,323 vectors.
  - Test Set: 83,099 vectors.
- Non-MIRS filtered speech (Female+Male)
  - Training Set: 101,359 vectors.
  - Test Set: 95,446 vectors.
- The magnitude was not quantized.

The segmental weighted signal-to-noise ratio (SNR) of the quantizer is illustrated in Figure 2. The proposed system achieves approximately 14dB SNR for as low as 6 bits for non-MIRS filtered speech, and nearly 10dB for MIRS [[20]] filtered speech.

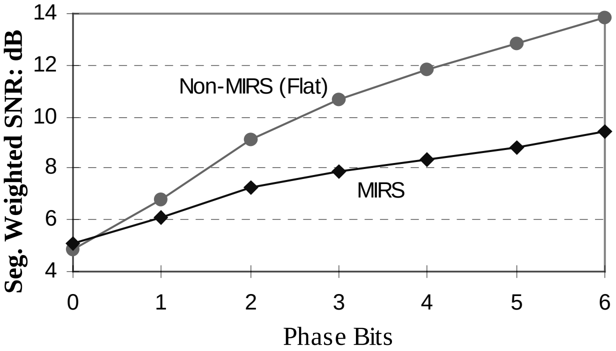

Figure 2. Segmental weighted SNR of the phase VQ versus the number of bits, for MIRS and for Non-MIRS (Flat) speech.

3.2. Subjective Results

Recent WI coders have used a male speaker extracted dispersion phase [[10],[15]]. We have conducted a subjective A/B test to compare our dispersion phase VQ, using only 4 bits, to a male extracted dispersion phase. The test data included 16 MIRS speech sentences, 8 of which are of female speakers, and 8 of male speakers. During the test, all pairs of file were played twice in alternating order, and the listeners could vote for either of the systems, or for no preference. The speech material was synthesized using WI system in which only the dispersion phase was quantized every 20ms. Twenty one listeners participated in the test. The test results, illustrated in Figure 3, show improvement in speech quality by using the 4-bit phase VQ. The improvement is larger for female speakers than for male. This may be explained by a higher number of bits per vector sample for female, by less spectral masking for female's speech, and by a larger amount of phase-dispersion variation for female.

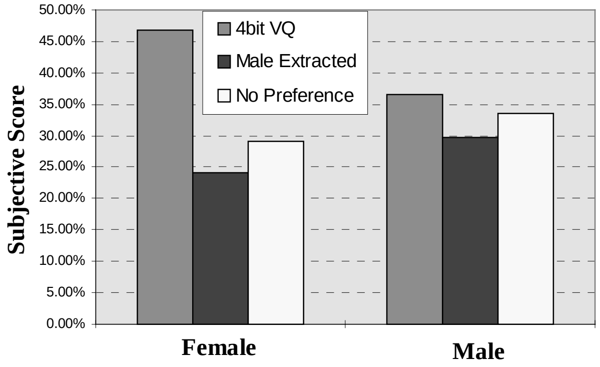

Figure 3. Results of subjective A/B test for comparison between the 4-bit phase VQ, and male extracted fixed phase.

The codebook design for the dispersion-phase quantization involves a tradeoff between robustness in terms of smooth phase variations and waveform matching. Locally optimized codebook for each pitch value may improve the waveform matching on the average, but may occasionally yield abrupt and excessive changes which may cause temporal artifacts.

4. SUMMARY

This paper presented an efficient Analysis-by-Synthesis vector quantizer for the dispersion phase of the excitation signal which was used to enhance a WI coder. The scheme incorporates perceptual weighting, and does not require any phase unwarping. The proposed quantizer achieves a segmental SNR of up to 14dB for as low as 6-bit quantization. Subjective testing shows improvement in synthesized speech quality using the quantized phase over a male speaker extracted phase. The improvement was larger for female speakers.

5. REFERENCES

[1] B. S. Atal, and M. R. Schroeder, “Stochastic Coding of Speech at Very Low Bit Rate”, *Proc. Int. Conf. Comm, Amsterdam,* pp. 1610-1613, 1984.

[2] 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, 1986.

[3] R. J. McAulay, and T. F. Quatiery, "Sinusoidal Coding", *in Speech Coding Synthesis by W. B. Kleijn and K. K. Paliwal, Elsevier Science B. V.*, Chapter 4, pp. 121-172, 1995.

[4] D. Griffin, and J. S. Lim, “Multiband Excitation Vocoder”, *IEEE Trans. ASSP*, Vol. 36, No. 8, pp. 1223-1235, August 1988.

[5] Y. Shoham, "High Quality Speech Coding at 2.4 to 4.0 kbps Based on Time-Frequency-Interpolation", *IEEE ICASSP'93*, Vol. II, pp. 167-170, 1993.

[6] W. B. Kleijn, "Encoding Speech Using Prototype Waveforms", *IEEE Trans. Speech and Audio Processing*, Vol. 1, No. 4, pp. 386-399, October 1993.

[7] W. B. Kleijn, and J. Haagen, "Transformation and Decomposition of The Speech Signal for Coding", *IEEE Signal Processing Letters*, Vol. 1, No. 9, pp. 136-138, 1994.

[8] W. B. Kleijn, and J. Haagen, "Speech Coder Based on Decomposition of Characteristic Waveforms", *IEEE ICASSP'95*, pp. 508-511, 1995.

[9] W. B. Kleijn, Y. Shoham, D. Sen, and R. Haagen, "A Low-Complexity Waveform Interpolation Coder", *IEEE ICASSP'96*, pp. 212-215, 1996.

[10] W. B. Kleijn, and J. Haagen, "Waveform Interpolation for Coding and Synthesis", *in Speech Coding Synthesis by W. B. Kleijn and K. K. Paliwal, Elsevier Science B. V.*, Chapter 5, pp. 175-207, 1995.

[11] I.S. Burnett, and R. J. Holbeche, "A Mixed Prototype Waveform/Celp Coder for Sub 3kb/p", *IEEE ICASSP'93*, Vol. II, pp. 175-178, 1993

[12] I. S. Burnett, and G. J. Bradley, "New Techniques for Multi-Prototype Waveform Coding at 2.84kb/s", *IEEE ICASSP'95*, pp. 261-263, 1995.

[13] I. S. Burnett, and G. J. Bradley, "Low Complexity Decomposition and Coding of Prototype Waveforms", *IEEE Speech Workshop*, pp. 23-24, 1995.

[14] I. S. Burnett, and D. H. Pham, "Multi-Prototype Waveform Coding using Frame-by-Frame Analysis-by-Synthesis", *IEEE ICASSP'97*, pp. 1567-1570, 1997.

[15] Y. Shoham, “Very Low Complexity Interpolative Speech Coding at 1.2 to 2.4 KBPS”, *IEEE ICASSP’97*, pp. 1599-1602, 1997.

[16] Y. Jiang, and V. Cuperman, “Encoding Prototype Waveforms Using A Phase Codebook”, *IEEE Workshop on Speech Coding for Telecommunications*, pp. 21-22, 1995

[17] W. R. Gardner, and B. D. Rao, “Noncausal All-Pole Modeling of Voiced Speech”, *IEEE Trans. Speech and Audio Processing*, Vol. 5, No. 1, pp.1-10, January 1997.

[18] X. Sun, et. al., “Phase Modeling of Speech Excitation for Low Bit-Rate Sinusoidal Transform Coding”, *IEEE ICASSP’97*, pp. 1691-1694, 1997.

[19] A. Gersho, and R. Gray, “Vector Quantization and Signal Compression”, *Kluwer Academic Publishers*, 1992.

[20] ITU-T, “Recommendation P.830, Subjective Performance Assessment of Telephone Band and Wideband Digital Codecs”, Annex D, *ITU*, Geneva, February 1996.