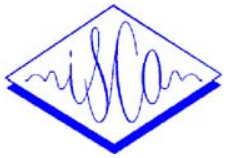

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ENHANCED ANALYSIS-BY-SYNTHESIS WAVEFORM INTERPOLATIVE CODING AT 4 KBPS

Oded Gottesman and Allen Gersho

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

ABSTRACT

This paper presents an *Enhanced analysis-by-synthesis* (AbS) *Waveform Interpolative* (EWI) speech coder at 4 kbps. The system incorporates novel features such as: AbS quantization of the *slowly evolving waveform* (SEW), AbS vector quantization (VQ) of the dispersion phase, a special pitch search for transitions, and switched-predictive analysis-by-synthesis gain VQ. Subjective quality tests indicate that it exceeds MPEG-4 at 4 kbps and of G.723.1 at 5.3 kbps, and it is slightly better than G.723.1 at 6.3 kbps.

1. INTRODUCTION

Recently, there has been growing interest in developing toll-quality speech coders at rates of 4 kbps and below. The speech quality produced by waveform coders such as *code-excited linear prediction* (CELP) coders [1] degrades rapidly at rates below 5 kbps. On the other hand, parametric coders such as the *waveform-interpolative* (WI) coder [4]-[6], the *sinusoidal-transform coder* (STC) [2], and the *multiband-excitation* (MBE) coder [3] produce good quality at low rates, but they do not achieve toll quality. This is mainly due to lack of robustness to parameter estimation, which is commonly done in open loop, and to inadequate modeling of non-stationary speech segments. In this work we propose a paradigm which incorporates AbS for parameter estimation, and a novel pitch search technique that is well suited for the non-stationary segments.

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 [4]-[6] typically use a fixed phase vector for the SEW, for example, in [5], a 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. Recently [8], we proposed a novel, efficient AbS VQ encoding of the dispersion phase of the excitation signal to enhance the performance of the WI coder at a very low bit-rate, which can be used for parametric coders as well as for waveform coders. The EWI coder employs this scheme, which incorporates perceptual weighting and does not require any phase unwrapping.

The WI coders use non-ideal low-pass filters for downsampling and upsampling of the SEW. We describe a novel AbS SEW quantization scheme, which takes the non-ideal filters into consideration. An improved match between reconstructed and original SEW is obtained, most notably in the transitions.

Pitch accuracy is crucial for high quality reproduced speech in WI coders. We introduce a novel pitch search technique based on varying segment boundaries; it allows for locking onto the most probable pitch period during transitions or other segments with rapidly varying pitch.

Commonly in speech coding the gain sequence is downsampled and interpolated. As a result it is often smeared during plosives and onsets. To alleviate this problem, we propose a novel switched-predictive AbS gain VQ scheme based on temporal weighting.

This paper is organized as follows. In Section 2 we explain the AbS SEW optimization. The dispersion phase quantizer is discussed in Section 3. Section 4 describes the pitch search. In Section 5 we present the switched-predictive AbS gain VQ. The bit allocation is given in section 6. Subjective results are reported in Section 7. Finally, we summarize our work.

2. AbS SEW QUANTIZATION

Commonly in WI coders the SEW is distorted by downsampling and upsampling with non-ideal low-pass filters. In order to reduce such distortion, an AbS SEW quantization scheme, illustrated in Figure 1, was used. Consider the accumulated weighted distortion, D_{wI}, between the input SEW vectors, $\mathbf{r}_m$, and the interpolated vectors, $\tilde{\mathbf{r}}_m$, given by:

$$D_{wI}(\hat{\mathbf{r}}_M, \{\mathbf{r}_m\}_{m=1}^{M+L-1}) = \left[\begin{array}{l} \sum_{m=1}^{M} [\mathbf{r}_m - \tilde{\mathbf{r}}_m]^H \mathbf{W}_m [\mathbf{r}_m - \tilde{\mathbf{r}}_m] \\ + \sum_{m=M+1}^{M+L-1} [1-\alpha(t_m)]^2 [\mathbf{r}_m - \tilde{\mathbf{r}}_M]^H \mathbf{W}_m [\mathbf{r}_m - \tilde{\mathbf{r}}_M] \end{array} \right] \quad (1)$$

where M is the number of waveforms per frame, L is the lookahead number of waveforms, $\alpha(t)$ is some increasing interpolation function in the range $0 \le \alpha(t) \le 1$, and $\mathbf{W}_m$ is a

This work was supported in part by the University of California MICRO program, ACT Networks, Inc, Cisco Systems, Inc., Conexant Systems, Inc., Dialogic Corp., DSP Group, Inc., Fujitsu Laboratories of America, Inc., General Electric Corp., Hughes Network Systems, Intel Corp., Lernout & Hauspie Speech Products NV, Lucent Technologies, Inc., Nokia Mobile Phones, Panasonic Speech Technology Laboratory, Qualcomm, Inc., Sun Microsystems Inc., and Texas Instruments, Inc.

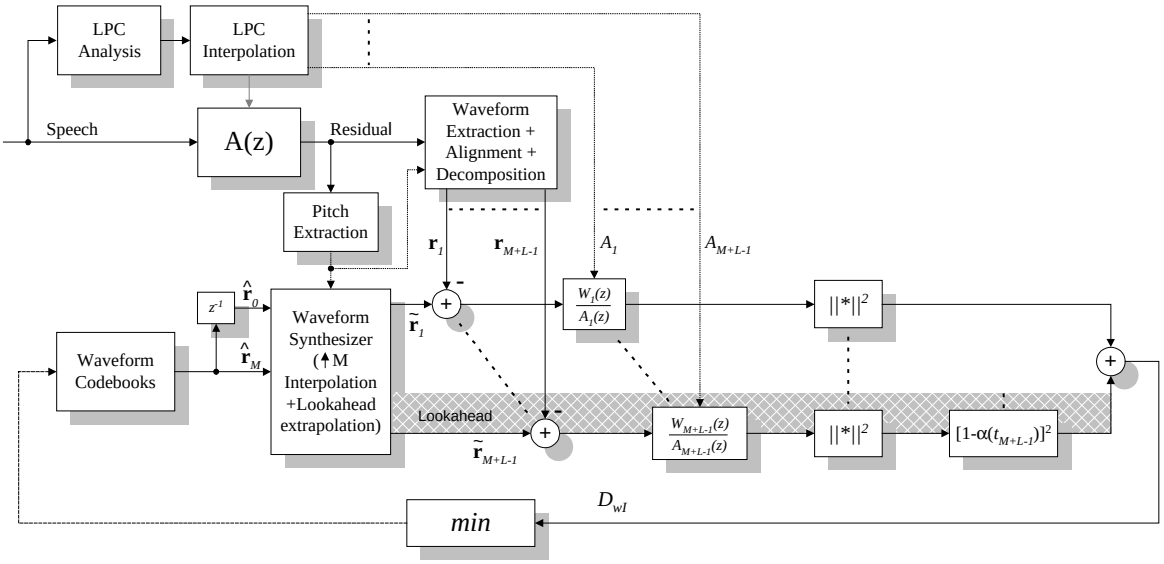

Figure 1. Block diagram of the AbS SEW vector quantization.

diagonal matrix whose elements, w_{kk}, are the combined spectral-weighting and synthesis of the k-th harmonic given by:

$$w_{kk} = \frac{1}{K}\left|\frac{gA(z/\gamma_1)}{\hat{A}(z)A(z/\gamma_2)}\right|^2_{z=e^{j(\frac{2\pi}{P})k}} \quad ; \; k = 1,..,K \qquad (2)$$

where P is the pitch period, K is the number of harmonics, g is the gain, $A(z)$ and $\hat{A}(z)$ are the input and the quantized LPC polynomials respectively, and the spectral weighting parameters satisfy $0 \le \gamma_2 < \gamma_1 \le 1$. The interpolated SEW vectors are given by:

$$\tilde{\mathbf{r}}_m = [1-\alpha(t_m)]\hat{\mathbf{r}}_0 + \alpha(t_m)\hat{\mathbf{r}}_M \quad ; \quad m = 1,..,M \qquad (3)$$

where, $\hat{\mathbf{r}}_0$ and $\hat{\mathbf{r}}_M$ are the quantized SEW at the previous and at the current frame respectively. It can be shown that the accumulated distortion in equation (1) is equal to the sum of *modeling distortion* and *quantization distortion*:

$$D_{wI}(\hat{\mathbf{r}}_M, \{\mathbf{r}_m\}_{m=1}^{M+L-1}) = D_{wI}(\mathbf{r}_{M,opt}, \{\mathbf{r}_m\}_{m=1}^{M+L-1}) + D_w(\hat{\mathbf{r}}_M, \mathbf{r}_{M,opt}) \qquad (4)$$

where the quantization distortion is given by:

$$D_w(\hat{\mathbf{r}}_M, \mathbf{r}_{M,opt}) = (\hat{\mathbf{r}}_M - \mathbf{r}_{M,opt})^H \mathbf{W}_{M,opt}(\hat{\mathbf{r}}_M - \mathbf{r}_{M,opt}) \qquad (5)$$

The optimal vector, $\mathbf{r}_{M,opt}$, which minimizes the modeling distortion, is given by:

$$\mathbf{r}_{M,opt} = \mathbf{W}_{M,opt}^{-1}\left[\sum_{m=1}^{M}\alpha(t_m)\mathbf{W}_m[\mathbf{r}_m - [1-\alpha(t_m)]\hat{\mathbf{r}}_0] + \sum_{m=M+1}^{M+L-1}[1-\alpha(t_m)]^2\mathbf{W}_m\mathbf{r}_m\right] \qquad (6)$$

where, $$\mathbf{W}_{M,opt} = \sum_{m=1}^{M}\alpha(t_m)^2\mathbf{W}_m + \sum_{m=M+1}^{M+L-1}[1-\alpha(t_m)]^2\mathbf{W}_m \qquad (7)$$

Therefore, VQ with the accumulated distortion of equation (1) can be simplified by using the distortion of equation (5), and:

$$\hat{\mathbf{r}}_M = \underset{\mathbf{r}'_i}{\operatorname{argmin}}\left\{ (\mathbf{r}'_i - \mathbf{r}_{M,opt})^H \mathbf{W}_{M,opt}(\mathbf{r}'_i - \mathbf{r}_{M,opt}) \right\} \qquad (8)$$

An improved match between reconstructed and original SEW is obtained, most notably in the transitions. Figure 2 illustrates the improved waveform matching obtained for a non-stationary speech segment by interpolating the optimized SEW.

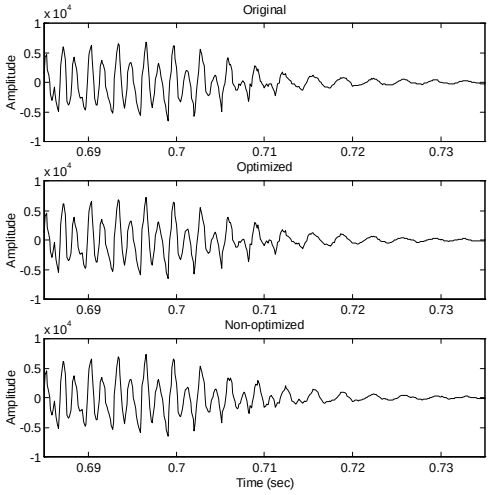

Figure 2. Example for the improved interpolation by SEW optimization during non stationary speech segment

3. AbS PHASE QUANTIZATION

The dispersion-phase quantization scheme [8][9] is illustrated in Figure 3. 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 domain by weighting it by the combined synthesis and weighting filter W(z)/A(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 and the quantized waveform. Phase dispersion quantization aims to improve *waveform matching*. Efficient AbS quantization can be obtained by using the perceptually weighted distortion:

$$D_w(\mathbf{r},\hat{\mathbf{r}}) = (\mathbf{r}-\hat{\mathbf{r}})^H \mathbf{W}(\mathbf{r}-\hat{\mathbf{r}}) \qquad (9)$$

The magnitude is perceptually more significant than the phase; and should therefore be quantized first. Furthermore, if the phase were quantized first, the very limited bit allocation available for the phase would lead to an excessively degraded spectral matching of the magnitude in favor of a somewhat improved, but less important, matching of the waveform. For the above distortion, the quantized phase vector is given by [8][9]:

$$\hat{\varphi} = \underset{\hat{\varphi}_i}{\operatorname{argmin}} \left\{ (\mathbf{r} - \mathbf{e}^{j\hat{\varphi}_i}|\hat{\mathbf{r}}|)^H \mathbf{W}(\mathbf{r} - \mathbf{e}^{j\hat{\varphi}_i}|\hat{\mathbf{r}}|) \right\} \qquad (10)$$

where i is the running phase codebook index, and $\mathbf{e}^{j\hat{\varphi}_i}$ is the respective diagonal phase exponent matrix. The AbS search for phase quantization is based on evaluating (10) for each candidate phase codevector. Since only trigonometric functions of the phase candidates are used, *phase unwrapping is avoided*. The EWI coder uses the optimized SEW, $\mathbf{r}_{M,opt}$, and the optimized weighting, $\mathbf{W}_{M,opt}$, for the AbS phase quantization.

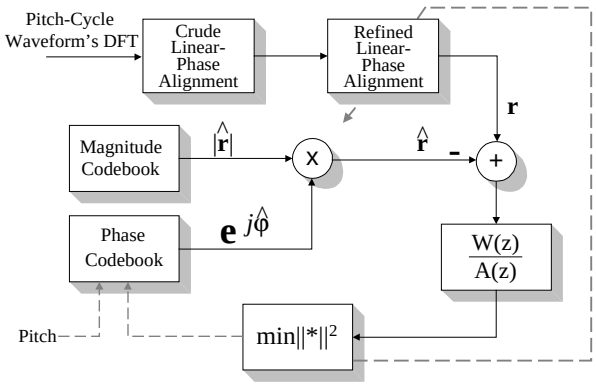

Figure 3. Block diagram of the AbS dispersion phase vector quantization.

4. PITCH SEARCH

The pitch search consists of a spectral domain search employed at 100 Hz and a temporal domain search employed at 500 Hz, as illustrated in Figure 4. The spectral domain pitch search is based on harmonic matching [2][3][7]. The temporal domain pitch search is based on varying segment boundaries. It allows for locking onto the most probable pitch period even during transitions or other segments with rapidly varying pitch. Initially, pitch periods, $P(n_i)$, are searched every 2 ms at instances n_i by maximizing the normalized correlation of the weighted speech $s_w(n)$, that is:

$$P(n_i) = \underset{\tau,N_1,N_2}{\arg\max} \left\{ \rho(n_i,\tau,N_1,N_2) \right\} = \qquad (11)$$

$$\underset{\tau,N_1,N_2}{\arg\max} \left\{ \frac{\sum_{n=n_i-N_1\Delta}^{n_i+\tau+N_2\Delta} s_w(n)s_w(n-\tau)}{\sqrt{\sum_{n=n_i-N_1\Delta}^{n_i+\tau+N_2\Delta} s_w(n)s_w(n)} \sqrt{\sum_{n=n_i-N_1\Delta}^{n_i+\tau+N_2\Delta} s_w(n-\tau)s_w(n-\tau)}} \right\}$$

where Δ is some incremental segment used in the summations for computational simplicity, and $0 \le N_j \le \lfloor 160/\Delta \rfloor$. Then, every 10 ms a weighted-mean pitch value is calculated by:

$$P_{mean} = \sum_{i=1}^{5} \rho(n_i)P(n_i) / \sum_{i=1}^{5} \rho(n_i) \qquad (12)$$

where $\rho(n_i)$ is the normalized correlation for $P(n_i)$.

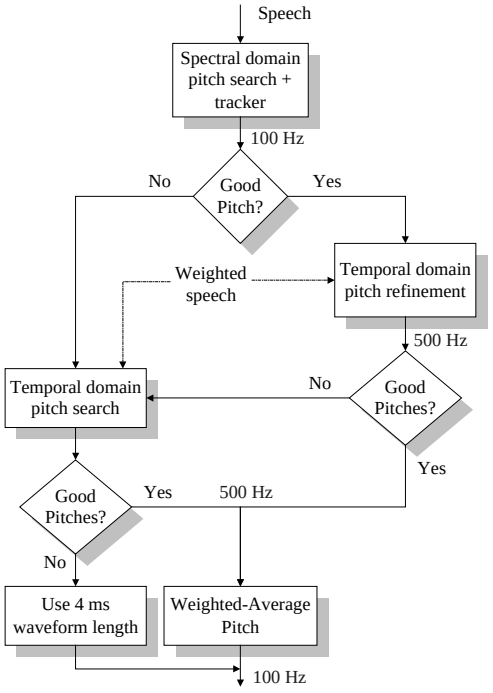

Figure 4. Pitch search of the EWI coder.

5. GAIN QUANTIZATION

The gain trajectory is commonly smeared during plosives and onsets by downsampling and interpolation. We address this problem and improve speech crispness with a novel Switched-Predictive AbS Gain VQ technique, illustrated in Figure 2. Switched-prediction is introduced to allow for different levels of gain correlation, and to reduce the occurrence of gain outliers. In order to improve speech crispness, especially for plosives and onsets, temporal weighting is incorporated in the AbS gain VQ. The weighting is a monotonic function of the temporal gain.

Two codebooks of 32 vectors each are used. Each codebook has an associated predictor coefficient, P_i, and a DC offset D_i. The quantization target vector is the DC removed log-gain vector denoted by *t(m)*. The search for the minimal WMSE is performed over all the vectors, $c_{ij}(m)$, of the codebooks. The quantized target, $\hat{t}(m)$, is obtained by passing the quantized vector, $c_{ij}(m)$, through the synthesis filter. Since each quantized target vector may have a different value of the removed DC, the quantized DC is added temporarily to the filter memory after the state update, and the next quantized vector's DC is subtracted from it before filtering is performed. Since the predictor coefficients are known, direct VQ can be used to simplify the computations.

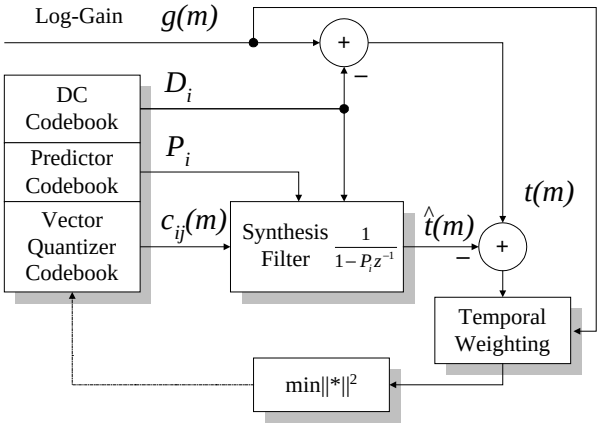

Figure 5. Switched-Predictive Analysis-by-Synthesis gain VQ using temporal weighting.

6. BIT ALLOCATION

The bit allocation of the coder is given in Table 1. The frame length is 20 ms, and ten waveforms are extracted per frame. The pitch and the gain are coded twice per frame.

Parameter	Bits / Frame	Bits / second
LPC	18	900
Pitch	2x6=12	600
Gain	2x6=12	600
REW	20	1000
SEW magnitude	14	700
SEW phase	4	200
Total	80	4000

Table 1. Bit allocation for EWI coder

7. SUBJECTIVE RESULTS

We have conducted a subjective A/B test to compare our 4 kbps EWI coder to MPEG-4 at 4 kbps, and to G.723.1. The test data included 24 MIRS speech sentences, 12 of which are of female speakers, and 12 of male speakers. Fourteen listeners participated in the test. The test results, listed in Table 2 to Table 4, indicate that the subjective quality of EWI exceeds that of MPEG-4 at 4 kbps and of G.723.1 at 5.3 kbps, and it is slightly better than that of G.723.1 at 6.3 kbps.

Test	4 kbps WI	4 kbps MPEG-4
Female	65.48%	34.52%
Male	61.90%	38.10%
Total	63.69%	36.31%

Table 2. Results of subjective A/B test for comparison between the 4 kbps WI coder to 4 kbps MPEG-4. With 95% certainty the WI preference lies in [58.63%, 68.75%].

Test	4 kbps WI	5.3 kbps G.723.1
Female	57.74%	42.26%
Male	61.31%	38.69%
Total	59.52%	40.48%

Table 3. Results of subjective A/B test for comparison between the 4 kbps WI coder to 5.3 kbps G.723.1. With 95% certainty the WI preference lies in [54.17%, 64.88%]

Test	4 kbps WI	6.3 kbps G.723.1
Female	54.76%	45.24%
Male	52.98%	47.02%
Total	53.87%	46.13%

Table 4. Results of subjective A/B test for comparison between the 4 kbps WI coder to 6.3 kbps G.723.1. With 95% certainty the WI preference lies in [48.51%, 59.23%].

8. SUMMARY

We have found several new techniques that enhance the performance of the WI coder. The most significant of these, reported here, *analysis-by-synthesis* vector-quantization of the dispersion-phase, AbS optimization of the SEW, a special pitch search for transitions, and switched-predictive analysis-by-synthesis gain VQ. These features improve the algorithm and its robustness. The test results indicate that the performance of the EWI coder slightly exceeds that of G.723.1 at 6.3 kbps and therefore EWI achieves very close to toll quality, at least under clean speech conditions.

9. 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, "Sinusoidal Coding", *in Speech Coding Synthesis by W. B. Kleijn and K. K. Paliwal, Elsevier Science B. V.*, Chapter 4, pp. 121-173, 1995.

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

[4] 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.

[5] 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.

[6] 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.

[7] E. Shlomot, V. Cuperman, and A. Gersho, "Hybrid Coding of Speech at 4 kbps", *IEEE Speech Coding Workshop*, pp. 37-38, 1997.

[8] O. Gottesman, "Dispersion Phase Vector Quantization For Enhancement of Waveform Interpolative Coder", *IEEE ICASSP'99*, vol. 1, pp. 269-272, 1999.

[9] O. Gottesman and A. Gersho, "Enhanced Waveform Interpolative Coding at 4 kbps", *IEEE Speech Coding Workshop*, 1999, Finland.