ENHANCING WAVEFORM INTERPOLATIVE CODING WITH WEIGHTED REW PARAMETRIC QUANTIZATION

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, Web: http://scl.ece.ucsb.edu

ABSTRACT

This paper presents an efficient quantization technique for the rapidly-evolving waveforms in *waveform interpolative* (WI) coders. The scheme, based on a parametrization of the *rapidly-evolving waveform* (REW) magnitude, and *analysis-by-synthesis* (AbS) *vector quantization* (VQ) of the REW parameters, allows both higher temporal and spectral resolution of the REW. A perceptually weighted distortion measure takes advantage of spectral and temporal masking and leads to improved reconstructed speech quality, most notably in mixed voiced and unvoiced speech segments. The technique is an important component of the *Enhanced Waveform Interpolative* (EWI) speech coder at 2.8 kbps that achieves a subjective quality slightly better than that of G.723.1 at 6.3 kbps.

1. INTRODUCTION

In WI coding [1]-[10], the similarity between successive REW magnitudes is generally exploited by downsampling and interpolation and by constrained bit allocation [1]. In our earlier EWI coder [7]-[9], the REW magnitude was quantized on a waveform by waveform basis, and with an excessive number of bits – more than is perceptually required. Here we propose a novel parametric representation of the REW magnitude and an efficient paradigm for AbS predictive vector quantization of the REW parameter sequence. The proposed scheme is discussed, and a simplified version is derived. The quantization scheme employs a perceptually weighted distortion measure, which takes advantage of spectral and temporal masking. The new method achieves a substantial reduction in the REW bit rate.

2. REW QUANTIZATION

Efficient REW quantization can benefit from two observations: (a) the REW magnitude is typically an increasing function of the frequency, which suggests that an efficient parametric representation may be used; (b) one can observe similarity between successive REW magnitude spectra, which suggests that employing predictive VQ on a group of adjacent REWs may yield useful coding gains. The next four sections introduce the REW parametric representation and the associated VQ technique.

2.1 REW Parameterization

Direct quantization of the REW magnitude is a variable dimension quantization problem, which may result in spending bits and computational effort on perceptually irrelevant information. A simple and practical way to obtain a reduced, and fixed dimension representation of the REW is with a linear combination of basis functions, such as orthonormal polynomial [4]-[6]. Such a representation usually produces a smoother REW magnitude, and improves the perceptual quality. Suppose the REW magnitude, $R(\omega)$, is represented by a linear combination of orthonormal functions, $\psi_i(\omega)$:

$$R(\omega) = \sum_{i=0}^{I-1} \gamma_i \psi_i(\omega) \quad , \quad 0 \le \omega \le \pi \tag{1}$$

where ω is the angular frequency, and I is the representation order. The REW magnitude is typically an increasing function of the frequency, which can be coarsely quantized with a small number of bits per waveform without significant perceptual degradation. Therefore, it may be advantageous to represent the REW magnitude in a simple, but perceptually relevant manner. Suppose the REW is modeled by the following parametric representation, $\hat{R}(\omega,\xi)$:

$$\hat{R}(\omega,\xi) = \sum_{i=0}^{I-1} \hat{\gamma}_i(\xi)\psi_i(\omega) \quad , \quad 0 \le \omega \le \pi \quad ; \quad 0 \le \xi \le 1 \tag{2}$$

where $\hat{\gamma}(\xi) = [\hat{\gamma}_0(\xi),...,\hat{\gamma}_{I-1}(\xi)]^T$ is a parametric vector of coefficients within the representation model subspace, and ξ is the "unvoicing" parameter which is zero for a fully voiced spectrum, and one for a fully unvoiced spectrum.

2.2 Piecewise Linear REW Representation

For practical considerations we may assume that the parametric representation is piecewise linear, and may be represented by a set of N uniformly spaced spectra, $\{\hat{R}(\omega,\bar{\xi}_n)\}_{n=0}^{N-1}$, as illustrated in Figure 1. This representation is similar to the hand-tuned REW codebook in [5], [6]. The parametric surface is linearly interpolated by:

$$\hat{R}(\omega,\xi) = (1-\alpha)\hat{R}(\omega,\bar{\xi}_{n-1}) + \alpha\hat{R}(\omega,\bar{\xi}_n) \tag{3}$$

$$; \quad \bar{\xi}_{n-1} \le \xi \le \bar{\xi}_n \quad ; \quad \alpha = \frac{\xi - \bar{\xi}_{n-1}}{\Delta\bar{\xi}} \quad ; \quad \Delta\bar{\xi} = \bar{\xi}_n - \bar{\xi}_{n-1}$$

From the linearity of the representation:

$$\hat{\gamma}(\xi) = (1-\alpha)\hat{\gamma}_{n-1} + \alpha\hat{\gamma}_n \tag{4}$$

where $\hat{\gamma}_n$ is the coefficient vector of the n-th REW magnitude representation, i.e. $\hat{\gamma}_n = \hat{\gamma}(\bar{\xi}_n)$.

2.3 REW Modeling

2.3.1 Non-Weighted Distortion

Suppose for a REW magnitude, $R(\omega)$, represented by some coefficient vector, γ, we search for the parameter value, $\xi(\gamma)$, in $\bar{\xi}_{n-1} \le \xi \le \bar{\xi}_n$, whose respective representation vector, $\hat{\gamma}(\xi)$, minimizes the MSE distortion between the two spectra:

$$D\left(R,\hat{R}(\xi)\right) = \int_0^{\pi} \left| R(\omega) - (1-\alpha)\hat{R}(\omega,\bar{\xi}_{n-1}) - \alpha\hat{R}(\omega,\bar{\xi}_n) \right|^2 d\omega \tag{5}$$

From orthonormality, the distortion is equal to:

$$D\left(R,\hat{R}(\xi)\right) = \left\|\gamma - \hat{\gamma}(\xi)\right\|^2 = \left\|\gamma - (1-\alpha)\hat{\gamma}_{n-1} - \alpha\hat{\gamma}_n\right\|^2 \tag{6}$$

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

The optimal interpolation factor that minimizes the MSE is:

$$\alpha_{opt} = \frac{(\hat{\gamma}_n - \hat{\gamma}_{n-1})^T (\gamma - \hat{\gamma}_{n-1})}{\left\| \hat{\gamma}_n - \hat{\gamma}_{n-1} \right\|^2} \tag{7}$$

and the respective optimal parameter value, which is a continuous variable between zero and one, is given by:

$$\xi(\gamma) = (1 - \alpha_{opt})\bar{\xi}_{n-1} + \alpha_{opt}\bar{\xi}_n \tag{8}$$

This result allows a rapid search for the best unvoicing parameter value needed to transform the coefficient vector to a scalar parameter, for encoding or for VQ design.

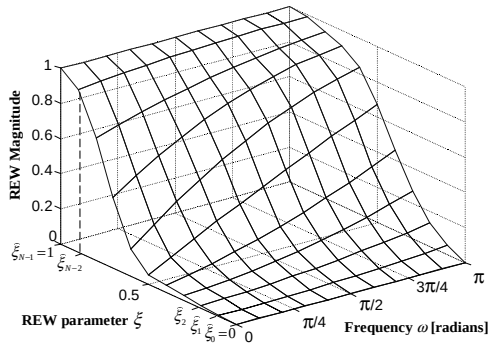

Fig. 1. REW Parametric Representation $\bar{R}(\omega, \xi)$

2.3.2 Weighted Distortion

Commonly in speech coding, quantization is performed with a perceptually weighted distortion measure. In this case, the weighted distortion between the input and the parametric representation modeled spectra is equal:

$$D_w\left(R, \hat{R}(\xi)\right) = \int_0^{\pi} \left| R(\omega) - \hat{R}(\omega, \xi) \right|^2 W(\omega) d\omega \tag{9}$$

$$= \left(\gamma - \hat{\gamma}(\xi)\right)^T \Psi\left(W(\omega)\right)\left(\gamma - \hat{\gamma}(\xi)\right)$$

where $\Psi\left(W(\omega)\right)$ is the weighted correlation matrix of the orthonormal functions, its elements are:

$$\Psi_{i,j}\left(W(\omega)\right) = \int_0^{\pi} W(\omega)\psi_i(\omega)\psi_j(\omega) d\omega \quad , \tag{10}$$

γ is the input coefficient vectors, and $\hat{\gamma}(\xi)$ is the modeled parametric coefficient vector. The optimal parameter that minimizes (9) is given by:

$$\alpha_{opt} = \frac{(\hat{\gamma}_n - \hat{\gamma}_{n-1})^T \Psi (\gamma - \hat{\gamma}_{n-1})}{(\hat{\gamma}_n - \hat{\gamma}_{n-1})^T \Psi (\hat{\gamma}_n - \hat{\gamma}_{n-1})} \tag{11}$$

and the respective optimal parameter value is computed using (8).

2.4 REW Quantization

2.4.1 Full Complexity Spectral Quantization Scheme

A novel switched-predictive AbS REW parameter VQ paradigm is illustrated in Fig. 2. Switched-prediction is introduced to allow for different levels of REW parameter correlation. The scheme incorporates both spectral weighting and temporal weighting. The spectral weighting is used for the distortion between each pair of input and quantized spectra. In order to improve SEW/REW mixing, particularly in mixed voiced and unvoiced speech segments, and to increase speech crispness, especially for plosives and onsets, temporal weighting is incorporated in the AbS REW VQ. The temporal weighting is a monotonic function of the temporal gain. A codebook with two partitions is used. Each partition has a particular predictor coefficient value, P_i, i=1 or 2. The quantization target is an *M*-dimensional vector of REW spectra. Each REW spectrum is represented by a vector of basis function coefficients denoted by $\gamma(m)$. The search for the minimal WMSE is performed over all the vectors, $\hat{c}_{ij}(m)$, of the codebook. The quantized REW function coefficients vector, $\hat{\gamma}\left(\hat{\xi}(m)\right)$, is a function of the quantized parameter $\hat{\xi}(m)$, which is obtained by passing the quantized vector, $\hat{c}_{ij}(m)$, through the synthesis filter. The weighted distortion between each pair of input and quantized REW spectra is calculated. The total distortion is a temporally-weighted sum of the *M* spectrally weighted distortions. Since the predictor coefficients are known, direct VQ can be used to simplify the computations. For a piecewise linear parametric REW representation, a substantial simplification of the search computations may be obtained by interpolating the distortion between the representation spectra set.

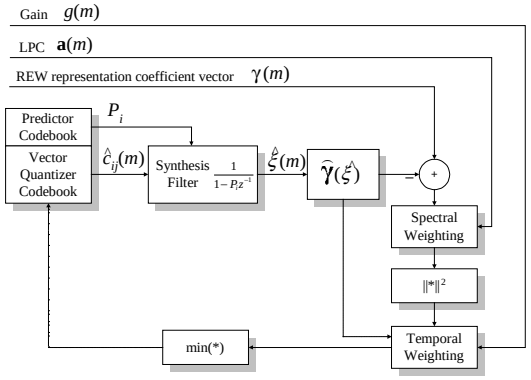

Fig. 2. REW Parametric Representation AbS VQ

2.4.2 Simplified Parametric Quantization Scheme

The above scheme maps each quantized parameter to a coefficient vector, which is used to compute the spectral distortion. To reduce complexity, such a mapping, and spectral distortion computation may be eliminated by using the simplified scheme described below. For high rate, and smooth representation function, the total distortion is equal the sum of modeling distortion and quantization distortion:

$$\sum_{m=1}^{M} D_w\left(R(m), \hat{R}(\hat{\xi}(m))\right) \tag{12}$$

$$= \sum_{m=1}^{M} D_w\left(R(m), \hat{R}(\xi(m))\right) + \sum_{m=1}^{M} D_w\left(\hat{R}(\xi(m)), \hat{R}(\hat{\xi}(m))\right)$$

The quantization distortion is related to the quantized parameter by:

$$\sum_{m=1}^{M} D_w\left(\hat{R}(\xi(m)), \hat{R}(\hat{\xi}(m))\right) \tag{13}$$

$$= \sum_{m=1}^{M} \left(\hat{\gamma}(\xi(m)) - \hat{\gamma}(\hat{\xi}(m))\right)^T \Psi\left(W(m)\right)\left(\hat{\gamma}(\xi(m)) - \hat{\gamma}(\hat{\xi}(m))\right)$$

which, for the piecewise linear representation case, is equal

$$\sum_{m=1}^{M} D_w\left(\hat{R}(\xi(m)), \hat{R}(\hat{\xi}(m))\right) \tag{14}$$

$$= \frac{1}{\Delta\hat{\xi}^2} \sum_{m=1}^{M} \Delta\hat{\gamma}_n(\xi(m))^T \Psi\left(W(m)\right) \Delta\hat{\gamma}_n(\xi(m))(\xi(m) - \hat{\xi}(m))^2$$

where $\Delta\hat{\gamma}_n(\xi(m)) = \hat{\gamma}_n(\xi(m)) - \hat{\gamma}_{n-1}(\xi(m))$. The quantization distortion is linearly related to the REW parameter squared quantization error, $\left(\xi(m) - \hat{\xi}(m)\right)^2$, and therefore justifies direct VQ of the REW parameter.

2.4.2.1 Simplified Scheme, Non-Weighted Distortion

The encoder maps the REW magnitude to an unvoicing parameter, and then quantizes the parameter by AbS VQ, as illustrated in Fig. 3. This scheme allows for higher temporal as well as spectral REW resolution, since no downsampling is performed, and the continuous parameter is vector quantized in AbS [10].

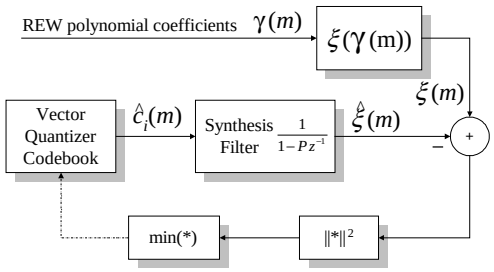

Fig. 3. REW Parametric Representation AbS VQ.

2.4.2.2 Simplified Scheme, Weighted Distortion

We may improve the quantization scheme to incorporate spectral and temporal weightings, as illustrated in Fig. 4. The REW parameter vector is first mapped to REW parameter by minimizing a distortion, which is weighted by the coefficient spectral weighting matrix $\mathbf{\Psi}$, as described in section 2.3.2. Then, the resulted REW parameter is used to compute a weighting, $w_s(\xi(m))$, in form of the spectral sensitivity to the REW parameter squared quantization error, $\left(\xi(m)-\hat{\xi}(m)\right)^2$, given by:

$$w_s(\xi(m)) = \left(\frac{\partial\hat{\gamma}}{\partial\xi}\right)^T \mathbf{\Psi}\left(\frac{\partial\hat{\gamma}}{\partial\xi}\right)_{\xi(m)} \tag{15}$$

For the piecewise linear representation case it is equal:

$$w_s(\xi(m)) = \frac{1}{\Delta\hat{\xi}^2}\Delta\hat{\gamma}_n(\xi(m))^T \mathbf{\Psi}(W(m))\Delta\hat{\gamma}_n(\xi(m)) \tag{16}$$

The above derivative can be computed off line. Additionally, a temporal weighting, in the form of a monotonic function of the gain, may be used, and denoted by $w_t(g(m))$. The AbS REW parameter quantization is computed by minimizing the combined spectrally and temporally weighted distortion:

$$D(\{\xi(m)\}_{m=1}^M, \{\hat{\xi}(m)\}_{m=1}^M) = \sum_{m=1}^M w_t(g(m))w_s(\xi(m))\left(\xi(m)-\hat{\xi}(m)\right)^2 \tag{17}$$

The weighted distortion scheme improves the reconstructed speech quality, most notably in mixed voiced and unvoiced speech segments. This may be explained by an improvement in REW/SEW mixing, and a less destructive REW contribution.

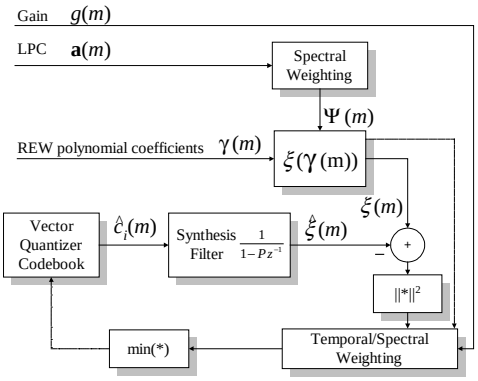

Fig. 4. REW Parametric Representation Simplified Weighted AbS VQ

3. BIT ALLOCATION

The bit allocation for the 2.8 kbps EWI coder is given in Table 1. The frame length is 20 ms, and ten waveforms are extracted per frame. The *line spectral frequencies* (LSFs) are coded using predictive MSVQ, having two stages of 10 bit each, a 2-bit increase compared to the past version of our coder [8], [9]. The pitch is coded twice per frame.

Parameter	Bits / Frame	Bits / second
LPC	20	1000
Pitch	2x6 = 12	600
Gain	9	450
SEW magnitude	8	400
REW magnitude	7	350
Total	56	2800

Table 1. Bit allocation for 2.8 kbps EWI coder

4. SUBJECTIVE RESULTS

We have conducted a subjective A/B test to compare our 2.8 kbps EWI coder to the G.723.1. The test data included 24 *modified intermediate reference system* (M-IRS) [11] filtered speech sentences, 12 of which are of female speakers, and 12 of male speakers. Twelve listeners participated in the test. The test results, listed in Table 2, indicate that the subjective quality of the 2.8 kbps EWI is slightly better than that of G.723.1 at 6.3 kbps.

Test	2.8 kbps WI	6.3 kbps G.723.1	No Preference
Female	38.19%	36.81%	25.00%
Male	43.06%	31.94%	25.00%
Total	40.63%	34.38%	25.00%

Table 2. Results of subjective A/B test for comparison between the 2.8 kbps EWI coder to 6.3 kbps G.723.1. With 95% certainty the result lies within +/-5.59%.

5. SUMMARY

We have found a new technique that enhances the performance of the WI coder, and allow for better coding efficiency. It offers an efficient parametrization of the REW magnitude, and AbS VQ of the REW parameter. This scheme allows for higher temporal as well as spectral REW resolution. The weighted distortion scheme improves the reconstructed speech quality, most notably in mixed voiced and unvoiced speech segments. Subjective test results indicate that the performance of the 2.8 kbps 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.

6. REFERENCES

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

[2] 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.

[3] 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.

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

[5] Y. Shoham, "Very Low Complexity Interpolative Speech Coding at 1.2 to 2.4 kbps," *IEEE ICASSP'97*, pp. 1599-1602, 1997.

[6] Y. Shoham, "Low-Complexity Speech Coding at 1.2 to 2.4 kbps Based on Waveform Interpolation," *International Journal of Speech Technology, Kluwer Academic Publishers*, pp. 329-341, May 1999.

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

[8] O. Gottesman and A. Gersho, "Enhanced Waveform Interpolative Coding at 4 kbps," *IEEE Speech Coding Workshop*, pp. 90-92, 1999, Finland.

[9] O. Gottesman and A. Gersho, "Enhanced Analysis-by-Synthesis Waveform Interpolative Coding at 4 kbps," EUROSPEECH'99, pp. 1443-1446, 1999, Hungary.

[10] O. Gottesman and A. Gersho, "High Quality Enhanced Waveform Interpolative Coding at 2.8 kbps," *IEEE ICASSP'00*, Turkey, 2000.

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