Enhanced Waveform Interpolative Coding at Low Bit-Rate

Oded Gottesman, *Member, IEEE,* and Allen Gersho, *Fellow, IEEE*

***Abstract*—This paper presents a high quality enhanced waveform interpolative (EWI) speech coder at low bit-rate. The system incorporates novel features such as optimization of the slowly evolving waveform (SEW) for interpolation, analysis-by-synthesis (AbS) vector quantization (VQ) of the SEW dispersion phase, dual-predictive AbS quantization of the SEW, efficient parameterization of the rapidly-evolving waveform (REW) magnitude, and VQ of the REW parameter, a special pitch search for transitions, and switched-predictive analysis-by-synthesis gain VQ. Subjective tests indicate that the 2.8 kb/s EWI coder's quality exceeds that of G.723.1 at 5.3 kb/s, and it is slightly better than that of G.723.1 at 6.3 kb/s.**

***Index Terms*—Analysis-by-synthesis, phase dispersion, speech coding, speech compression, vector quantization, waveform interpolation, waveform interpolative coding.**

I. Introduction

In RECENT years, there has been increasing interest in achieving toll-quality speech coding at rates of 4 kb/s and below. Currently, there is an ongoing 4 kb/s standardization effort conducted by the ITU-T. The expanding variety of emerging applications for speech coding, such as third generation wireless networks and Low Earth Orbit (LEO) systems, is motivating increased research efforts. The speech quality produced by waveform coders such as *code-excited linear prediction* (CELP) coders [1] degrades rapidly at rates below 5 kb/s. On the other hand, parametric coders such as the *waveform-interpolative* (WI) coder [8]–[20], the *sinusoidal-transform coder* (STC) [2], the *multiband-excitation* (MBE) coder [3], the *mixed-excitation linear predictive* (MELP) vocoder [4], [5], and the *harmonic-stochastic excitation* (HSX) coder [6] produce good quality at low rates, but they do not achieve toll quality. This is largely due to the lack of robustness of speech parameter estimation, which is commonly performed in open-loop, and to inadequate modeling of nonstationary speech segments. In this work we propose a paradigm for WI coding that incorporates *analysis-by-synthesis* (AbS) for parameter estimation, offers higher temporal and spectral resolution for the *rapidly-evolving waveform* (REW), and more efficient quantization of the *slowly-evolving waveform* (SEW).

The WI coders [13]–[20] use nonideal low-pass filters for downsampling and upsampling of the SEW. We describe a novel AbS SEW quantization scheme, which takes the nonideal filters into consideration. An improved match between reconstructed and original SEW spectra is obtained, most notably in transition segments of speech.

Commonly in WI coding, the similarity between successive REW magnitudes is exploited by downsampling and interpolation and by bit allocation that constrains similarity [13]. In our previous enhanced waveform-interpolative (EWI) coder [22], [23], 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 new method achieves a substantial reduction in the REW bit-rate.

In low bit-rate WI coding, the relation between the SEW and the REW magnitudes was exploited by computing the magnitude of one as the unity complement of the other [14], [17]–[20]. Also, since the sequence of SEW spectrum evolves slowly, successive SEWs exhibit similarity, offering opportunities for redundancy removal. Additional forms of redundancy that may be exploited for coding efficiency are 1) for a fixed SEW/REW decomposition filter, the mean SEW magnitude increases with the pitch period and 2) the similarity between successive SEWs, also increases with the pitch period. These phenomena are due to the fact that, for uniformly extracted waveforms, the overlap between successive waveforms increases with the pitch period. In this work, we introduce a novel "dual-predictive" AbS paradigm for quantizing the SEW magnitude that optimally exploits the information about the current quantized REW, the past quantized SEW, and the pitch, in order to estimate the current SEW.

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 [8]–[20] typically use a fixed phase vector for the SEW, for example, in [14], [19], 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. In the past [31]–[34], phase modeling and quantization

Manuscript received April 25, 2000; revised June 19, 2001. This work was supported in part by the National Science Foundation under Grant MIP-9707764, the University of California MICRO Program, Cisco Systems, Inc., Conexant Systems, Inc., Dialogic Corp., Fujitsu Laboratories of America, Inc., General Electric Co., Hughes Network Systems, Lernout & Hauspie Speech Products, Lockheed Martin, Lucent Technologies, Inc., Panasonic Speech Technology Laboratory, Qualcomm, Inc., and Texas Instruments, Inc. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Peter Kroon.

A. Gersho is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: gersho@ece.ucsb.edu; http://scl.ece.ucsb.edu).

O. Gottesman is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA and also with Compandent, Inc., Goleta, CA 93117 USA (e-mail: gottesman@ece.ucsb.edu; oded@gottesmans.com; http://www.compandent.com).

Publisher Item Identifier S 1063-6676(01)08235-9.

was investigated. In [32] a random phase codebook was used at a relatively high number of phase quantization bits. In [33], [34], a noncausal all-pole filter's phase model was discussed, but quantization was not optimized. We have observed that such a model is quite inadequate in matching the physiological excitation's phase, although occasionally it does provide a reasonable match. In addition, none of the above methods have incorporated perceptual weighting. Recently [21], 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 low bit-rate, which can be used for parametric coders as well as for waveform coders. The EWI coder presented here employs this scheme, which incorporates perceptual weighting and does not require any phase unwrapping.

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. In the past, this problem was addressed by employing a special mechanism that mimicked the gain characteristics [14]. To alleviate this problem, we propose a novel switched-predictive AbS gain VQ scheme based on temporal weighting.

This paper is organized as follows. Section II describes the WI coder. In Section III we explain the AbS SEW optimization. The dispersion phase quantizer is discussed in Section IV. In Section V we explain the REW parameterization, and the corresponding AbS VQ. The dual predictive SEW AbS VQ and its performance are discussed in Section VI. Section VII describes the pitch search. In Section VIII we present the switched-predictive AbS gain VQ. The bit allocation is given in Section IX. Subjective results are reported in Section X. Finally, we summarize our work.

II. Description of the Waveform Interpolative Coder

A. Introduction to Waveform Interpolation

During voiced speech, which is quasiperiodic, one can observe the underlying process of evolving shape of successive pitch cycles. A continuously evolving sequence of pitch cycle waveforms can be generated from a continuous-time signal, either from the linear prediction residual or from the speech waveform directly. For coding purposes, one may extract a subsequence of these waveforms, and apply quantization to it. At the decoder, following inverse quantization, speech synthesis can be performed by interpolating missing waveforms. Such a process is the essence of waveform interpolative coding [8]–[20].

Speech segments typically contain both voiced and unvoiced attributes. The different perceived character of the voiced and unvoiced components [27] suggests a separation of the components, and applying distinct perceptually based coding to them [12]–[20].

B. Definitions

Given a continuous linear prediction residual (or speech) signal, $e(t)$, and its associated instantaneous pitch period contour, $p(t)$, a *characteristic waveform* (CW) [8]–[20], $u(t,\phi)$, may be generated by extracting pitch cycles at an infinitely high rate, normalizing their length to 2π, and aligning them sequentially by a cyclical shift. The differential alignment phase shift, $d\phi_s$, is given by

$$d\phi_s = \frac{2\pi}{p(t)}dt. \tag{1}$$

Therefore, the temporal accumulated phase shift is equal to

$$\phi_s(t) = \phi_0 + \int_{t_0}^{t} \frac{2\pi}{p(t')}\,dt' \tag{2}$$

where ϕ_0 is the initial phase shift at time t_0. The CW is a two-dimensional (2-D) surface which is defined by

$$u(t,\phi) \equiv e\left(t + \frac{p(t)}{2\pi}[\phi - \phi_s(t)]_\pi\right) \tag{3}$$

where $[x]_\pi$ wraps x over the range $[-\pi,\pi)$, and is defined by

$$[x]_\pi \equiv ((x+\pi) \text{ modulo } 2\pi) - \pi. \tag{4}$$

The CW is a periodic function of the parameter ϕ, with a period 2π. The residual (or speech) signal may be generated from the CW by calculating its value along the phase shift contour

$$e(t) = u(t,\phi)_{|_{\phi=\phi_s(t)}} = u(t,\phi_s(t)). \tag{5}$$

The WI coder based on this 2-D function is conceptually similar to the pitch synchronous transform coder [7].

C. Waveform Interpolative Coder Description

The EWI coder is based on the WI coding model [11]–[14]. In this model, the CW is decomposed into two components called SEW and REW. The SEW, which is computed by low-pass filtering the 2-D CW surface along the time axis (also known as the evolutionary axis), contains most of the voiced speech attribute. The SEW is coded at low temporal resolution, high spectral resolution, and using spectrally weighted distortion measure. The REW, which is the complementary high-pass component, represents primarily the unvoiced speech attribute. The REW is coded at high temporal resolution, low spectral resolution, and by exploiting spectral and temporal masking.

The EWI encoder is illustrated in Fig. 1. The LPC analysis, and quantization is performed every 20 ms frame, and interpolated values are used for each of the ten waveforms in the frame. The input speech is then passed through the resulting whitening filter to produce the residual signal. A search for the pitch period is performed and the pitch is quantized every 10 ms, and is then interpolated. The interpolated pitch values are used for pitch cycle waveform extraction, which is performed at a regular rate (every 2 ms). The rate must be higher then the maximal pitch frequency in order to prevent aliasing along the time axis [14], [18]. The extracted waveforms are then power normalized, and sequentially aligned, to form a discrete-time CW, which is represented by a Fourier series (FS). The Fourier coefficients

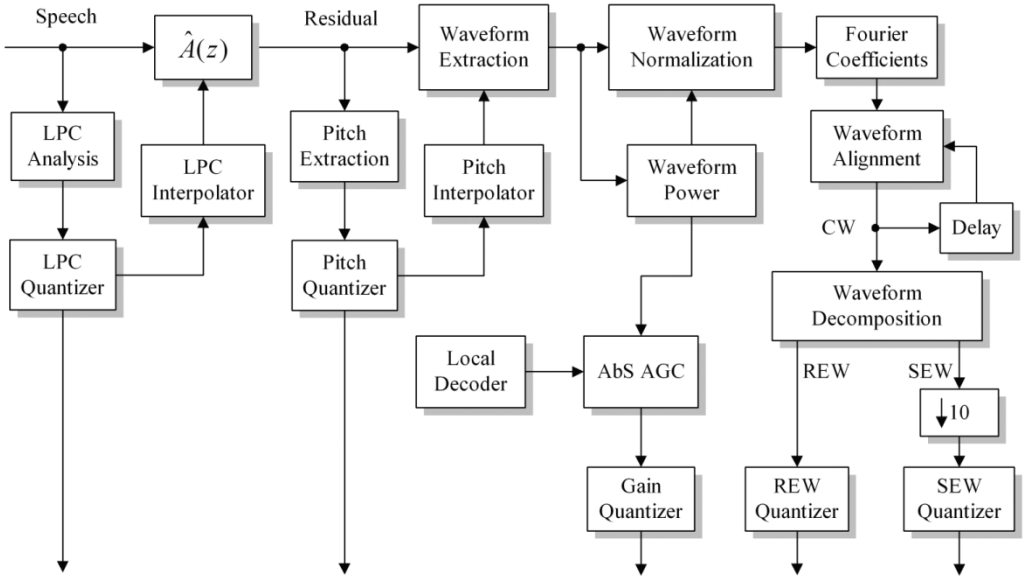


Fig. 1. Block diagram of the EWI encoder.

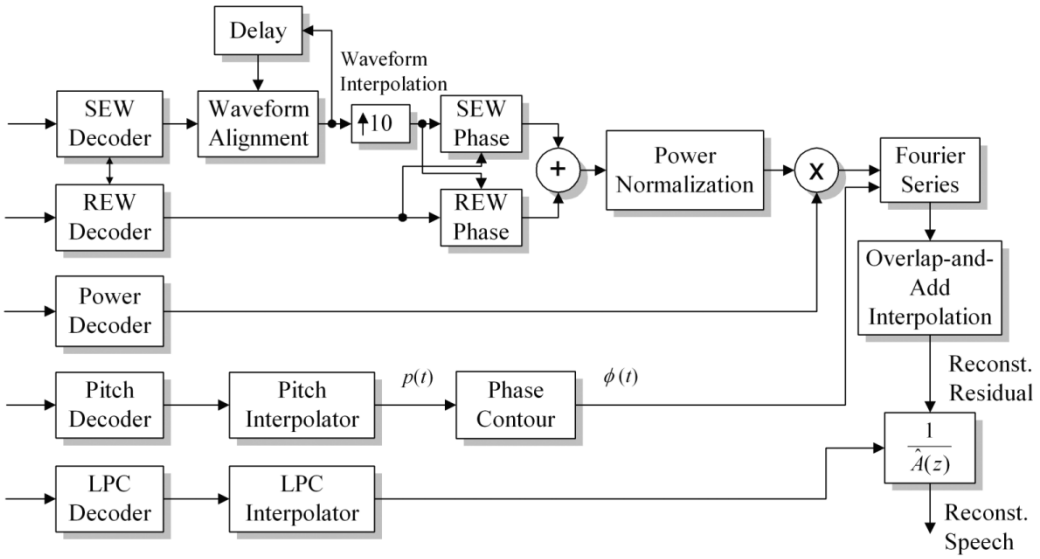


Fig. 2. Block diagram of the EWI decoder.

(FCs) are obtained by pitch-synchronous discrete Fourier transform (DFT). The frequency domain representation is used in order to benefit from appropriate perceptually motivated coding paradigms for the magnitude, and the phase. The CW is then low-pass filtered along the time axis, to produce the SEW. The REW is computed as the complementary high-pass component, and is then quantized. The SEW is downsampled, and then quantized every 20 ms. Finally, a local decoder is used to reconstruct the speech, then the encoder adjusts the gain to equate the reconstructed speech waveform energy to that of the input speech waveform, and quantizes the resultant gain.

The EWI decoder is illustrated in Fig. 2. The REW and the SEW are decoded, and an interpolated SEW is computed each 2 ms. The REW and SEW are phase adjusted to achieve adequate voicing level and to benefit from temporal masking, and then added together. The resulting waveform is then power-normalized, and multiplied by the respective quantized gain. The pitch is decoded, and interpolated, and is then used for computing the phase contour using (2). The reconstructed residual is computed by continuous waveform interpolation, which is performed by computing the Fourier series along the phase contour followed by overlap-and-add. Over the interpolation interval $t_m \leq t \leq t_{m+1}$, the continuous reconstructed excitation signal, $\hat{e}(t)$, is given by

$$\hat{e}(t) = [(1-\alpha(t))\hat{u}(t_m,\phi) + \alpha(t)\hat{u}(t_{m+1},\phi)]_{|\phi=\phi_s(t)} \quad (6)$$

where $\hat{u}(t_m,\phi)$ and $\hat{u}(t_{m+1},\phi)$ are the reconstructed CW at the interval beginning and ending, respectively, and $\alpha(t)$ is some increasing interpolation function in the range $0 \leq \alpha(t) \leq 1$. The quantized LPC coefficients are interpolated, and are then used for the synthesis filter. Finally, the reconstructed speech is obtained by passing the reconstructed residual through the syn-

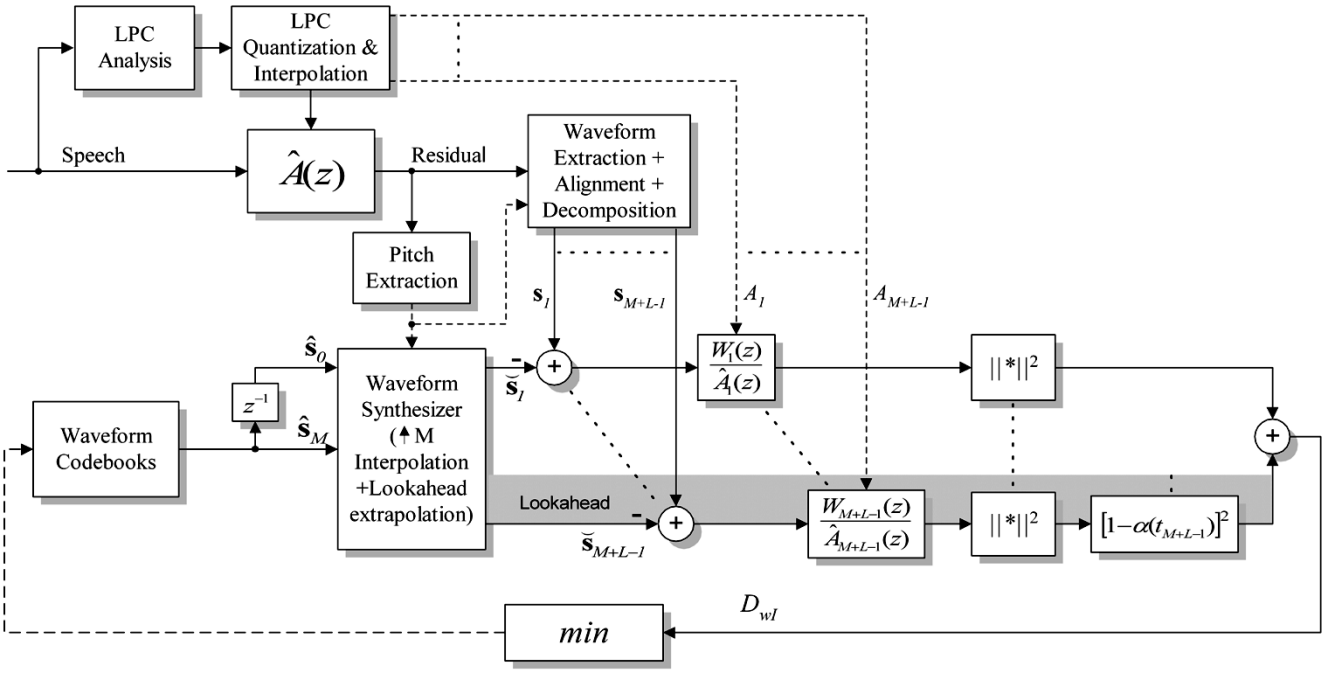


Fig. 3. Block diagram of the AbS SEW vector quantization.

thesis filter. For low rate coding, it is beneficial to use a formant adaptive postfilter [28]. In WI coding the postfilter enhances the quantized speech quality by reducing the audibility of the nonperiodic speech component around the formants. Such component is mostly due to the REW which is still somehow related to the SEW and may not always be regarded as independent noise.

Many speech coding schemes use voiced/unvoiced classification with separate coding of each type of sound. Such schemes may suffer severe quality loss whenever classification error is made, which causes the coder to apply coding method that is inappropriate to the coded speech sound. One of the important advantages of the WI coding system is that it is universally applied to all speech sounds, and is therefore more robust than classification based coding scheme.

III. SEW Optimization

Most WI coders [10]–[18] use nonideal low-pass filters for downsampling and upsampling of the SEW. These filters introduce aliasing and mirroring distortion, even when no quantization is applied. We propose, instead, a novel AbS SEW quantization scheme, illustrated in Fig. 3, which takes the nonideal interpolation filters into consideration and optimizes the SEW accordingly, however some aliasing may already exist (due to nonideal anti-aliasing filters) and this will not be eliminated by the AbS quantization scheme. The input speech is analyzed and LPC parameters are extracted, quantized and interpolated, and an LPC whitening filter is obtained. Then the speech is passed through the resulting whitening filter to produce the residual signal. In each frame M SEWs are extracted from the residual with L look-ahead waveforms. Each waveform is represented by a vector of FCs $\mathbf{s}_m$. The local decoder at the encoder reconstructs M SEWs, $\{\breve{\mathbf{s}}_m\}_{m=1}^{M}$, by interpolating between the

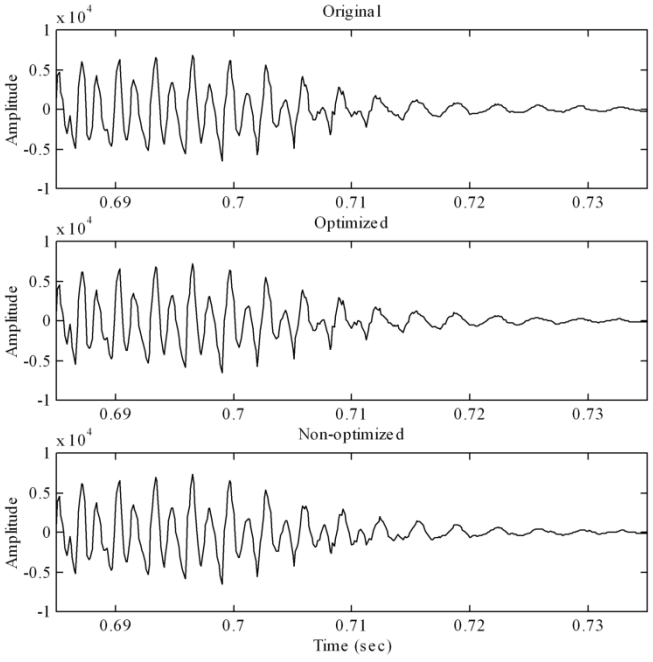


Fig. 4. Example for the improved interpolation by SEW optimization during nonstationary speech segment.

quantized SEW at the previous frame, $\hat{\mathbf{s}}_0$, to the current frame quantized SEW, $\hat{\mathbf{s}}_M$. The interpolated SEW vectors are given by

$$\breve{\mathbf{s}}_m = [1-\alpha(t_m)]\hat{\mathbf{s}}_0 + \alpha(t_m)\hat{\mathbf{s}}_M; \quad m = 1, \ldots, M. \tag{7}$$

Assuming $\hat{\mathbf{s}}_0$ and the LPC coefficients are given, the encoder's task is to find the quantized vector $\hat{\mathbf{s}}_M$ such that the *accumulated weighted distortion between original and reconstructed waveform sequences*, denoted by $D_{w\mathrm{I}}$, is minimized. Since the

effect of the linear interpolation LPF is taken into account in the proposed scheme, a true interpolated waveform (synthesis) is incorporated in the analysis process, unlike the conventional open-loop WI coders [10]–[18] in which only one waveform, namely $\mathbf{s}_M$, is used for the quantization. Consider the accumulated weighted distortion, $D_{w\mathrm{I}}$, between the input SEW FCs vectors, $\mathbf{s}_m$, and the quantized and interpolated vectors, $\breve{\mathbf{s}}_m$, given by

$$\begin{aligned} &D_{w\mathrm{I}}\left(\hat{\mathbf{s}}_M, \{\mathbf{s}_m\}_{m=1}^{M+L-1}\right) = \\ &\textstyle\sum_{m=1}^{M}[\mathbf{s}_m - \breve{\mathbf{s}}_m]^H \mathbf{W}_m [\mathbf{s}_m - \breve{\mathbf{s}}_m] \\ &\quad + \textstyle\sum_{m=M+1}^{M+L-1}[1-\alpha(t_m)]^2 [\mathbf{s}_m - \breve{\mathbf{s}}_M]^H \mathbf{W}_m [\mathbf{s}_m - \breve{\mathbf{s}}_M] \end{aligned} \tag{8}$$

where

M	number of waveforms per frame;
L	number of look-ahead waveforms;
$\mathbf{W}_m$	diagonal matrix whose elements, w_{kk}, are the spectral values of the combined spectral-weighting and synthesis filters at the kth harmonic given by

$$w_{kk} = \frac{1}{K}\left|\frac{gA(z/\gamma_1)}{\hat{A}(z)A(z/\gamma_2)}\right|^2_{z=e^{j(\frac{2\pi}{P})k}} \qquad k = 1, \ldots, K \tag{9}$$

where

P	pitch period;
K	number of harmonics;
g	gain;
$A(z)$ and $\hat{A}(z)$	input and the quantized LPC polynomials, respectively.

The spectral weighting parameters satisfy $0 \le \gamma_2 < \gamma_1 \le 1$. It can be shown that the accumulated distortion in (8) is equal to the sum of two components, a *modeling distortion* and a *quantization distortion*

$$\begin{aligned} D_{w\mathrm{I}}\left(\hat{\mathbf{s}}_M, \{\mathbf{s}_m\}_{m=1}^{M+L-1}\right) &= D_{w\mathrm{I}}\left(\mathbf{s}_{M,\mathrm{opt}}, \{\mathbf{s}_m\}_{m=1}^{M+L-1}\right) \\ &\quad + D_w(\hat{\mathbf{s}}_M, \mathbf{s}_{M,\mathrm{opt}}) \end{aligned} \tag{10}$$

where the quantization distortion is given by

$$D_w(\hat{\mathbf{s}}_M, \mathbf{s}_{M,\mathrm{opt}}) = (\hat{\mathbf{s}}_M - \mathbf{s}_{M,\mathrm{opt}})^H \mathbf{W}_{M,\mathrm{opt}} (\hat{\mathbf{s}}_M - \mathbf{s}_{M,\mathrm{opt}}) \tag{11}$$

where the optimal vector, $\mathbf{s}_{M,\mathrm{opt}}$, (which minimizes the modeling distortion) is given by

$$\mathbf{s}_{M,\mathrm{opt}} = \mathbf{W}_{M,\mathrm{opt}}^{-1} \begin{bmatrix} \sum_{m=1}^{M} \alpha(t_m)\mathbf{W}_m[\mathbf{s}_m - [1-\alpha(t_m)]\hat{\mathbf{s}}_0] \\ + \sum_{m=M+1}^{M+L-1}[1-\alpha(t_m)]^2 \mathbf{W}_m \mathbf{s}_m \end{bmatrix} \tag{12}$$

and the respective weighting matrix is given by

$$\mathbf{W}_{M,\mathrm{opt}} = \sum_{m=1}^{M} \alpha(t_m)^2 \mathbf{W}_m + \sum_{m=M+1}^{M+L-1} [1-\alpha(t_m)]^2 \mathbf{W}_m. \tag{13}$$

Therefore, VQ with the accumulated distortion of (8) can be simplified by using the distortion of (11), and

$$\hat{\mathbf{s}}_M = \underset{\mathbf{s}'_i}{\mathrm{argmin}}\{(\mathbf{s}'_i - \mathbf{s}_{M,\mathrm{opt}})^H \mathbf{W}_{M,\mathrm{opt}} (\mathbf{s}'_i - \mathbf{s}_{M,\mathrm{opt}})\}. \tag{14}$$

An improved match between reconstructed and original SEW is obtained, most notably in the transitions. Fig. 4 illustrates the improved waveform matching obtained for a nonstationary speech segment by interpolating the optimized SEW.

IV. Dispersion Phase Quantization

The dispersion-phase quantization scheme [21]–[23] is illustrated in Fig. 5. A pitch cycle that is extracted from the SEW is applied as an input to the system, and is cyclically shifted so that its pulse is located at position zero. Let its FC vector be denoted by $\mathbf{s}$. After quantization, the components of the quantized magnitude vector, $|\hat{\mathbf{s}}|$, are multiplied by the exponential of the quantized phases, $\hat{\varphi}_k$, to yield the quantized waveform FC vector, $\hat{\mathbf{s}}$, which is subtracted from the input FC vector to produce the error FC vector. The error FC vector is then transformed to the perceptually-weighted frequency domain by weighting it by the combined synthesis and weighting filter $W(z)/\hat{A}(z)$. The encoder searches for the phase that minimizes the energy of the perceptually weighted error, allowing a fine tuning of the 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

$$D_w = \frac{1}{2\pi}\int_0^{2\pi} |s_w(\phi) - \hat{s}_w(\phi)|^2 \, d\phi \tag{15}$$

where $s_w(\phi)$ is the weighted input SEW prototype and $\hat{s}_w(\phi)$ is the quantized and weighted SEW prototype. It can be shown [21] that the above distortion is equivalent to

$$D_w(\mathbf{s}, \hat{\mathbf{s}}) = (\mathbf{s} - \hat{\mathbf{s}})^H \mathbf{W} (\mathbf{s} - \hat{\mathbf{s}}). \tag{16}$$

The magnitude is perceptually more significant than the phase [26] 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 this distortion measure, the quantized phase vector is given by [21]–[23]

$$\hat{\boldsymbol{\varphi}} = \underset{\hat{\boldsymbol{\varphi}}_i}{\mathrm{argmin}}\{(\mathbf{s} - \mathbf{e}^{j\hat{\boldsymbol{\varphi}}_i}|)^H \mathbf{W} (\mathbf{s} - \mathbf{e}^{j\hat{\boldsymbol{\varphi}}_i}|\hat{\mathbf{s}}|)\} \tag{17}$$

where

i	running phase codebook index;
$\mathbf{e}^{j\hat{\boldsymbol{\varphi}}_i}$	respective diagonal phase exponent matrix;
$\lvert\hat{\mathbf{s}}\rvert$	quantized magnitude vector.

The AbS search for phase quantization is based on evaluating (17) for each candidate phase codevector. Since only trigonometric functions of the phase candidates are used (via complex exponentials), only phase values modulo 2π are relevant, and therefore *phase unwrapping is avoided*. The EWI coder uses the optimized SEW, $\mathbf{s}_{M,\mathrm{opt}}$, and the optimized weighting, $\mathbf{W}_{M,\mathrm{opt}}$, for the AbS phase quantization.

A. Phase Centroid Equations

We will now describe the training of the phase codebook. Suppose $\Im = \{\mathbf{S}_n; n = 1, 2, \ldots, \|\Im\|\}$ is the set of SEW

training vectors used for the design of the phase VQ, where $\|\Im\|$ is the *cardinality* of the set $\Im$, that is, the number of elements in $\Im$. The average global distortion measure for the quantization of the training set is

$$D_{w,\text{Global}} = \frac{1}{\|\Im\|}\sum_{n=1}^{\|\Im\|} D_w(\mathbf{S}_n, \mathbf{e}^{j\hat{\boldsymbol{\varphi}}_n}|\hat{\mathbf{S}}_n|) = \frac{1}{\|\Im\|}\sum_{n=1}^{\|\Im\|}\sum_{k=1}^{K_n} w_{kk,n}|S_{k,n} - e^{j\hat{\varphi}_{k,n}}|\hat{S}_{k,n}||^2 \quad (18)$$

where $S_{k,n}$, and $\hat{S}_{k,n}$ are the kth FC of the nth input and the quantized SEWs, respectively. The jth optimal partition cell satisfies

$$\Re_j = \{\mathbf{S} : D_w(\mathbf{S}, \mathbf{S}_j) \le D_w(\mathbf{S}, \hat{\mathbf{S}}_i);\ \text{for all } i\}. \quad (19)$$

For a given partition $\Re = \{\Re_j; j = 1, 2, \ldots, J\}$, the centroid equation [29] of the kth coefficient's phase, for the jth cluster, which minimizes the global distortion (18), is given by

$$\hat{\varphi}(k)_{j\text{th-cluster}} = \text{atan}\left[\frac{\sum_{\mathbf{S}_n\in\Re_j} w_{kk,n}|\hat{S}_{k,n}||S_{k,n}|\sin(\varphi_{k,n})}{\sum_{\mathbf{S}_n\in\Re_j} w_{kk,n}|\hat{S}_{k,n}||S_{k,n}|\cos(\varphi_{k,n})}\right]; \quad j = 1, \ldots, J. \quad (20)$$

B. Variable Dimension Vector Quantization

The phase vector's dimension depends on the pitch period and, therefore, a variable dimension VQ has been implemented. In our WI coder, the possible pitch period value was divided into several ranges, and for each range of pitch period a codebook was designed such that all vectors of dimension smaller than the largest pitch period in that range are zero padded beyond their highest element. Pitch changes over time cause the quantizer to switch among the pitch-range selected codebooks. In order to achieve smooth phase variations whenever such a switch occurs, overlapped training clusters were used and similar initial conditions were selected for each codebook. This design method does not guarantee smoothness, i.e., for a slight change in pitch that causes a switch in codebooks, the quantized vector could change substantially. However, significant quality improvement was obtained with the procedure. We believe such smoothness may be guaranteed by including some heuristic rules in the encoding process.

C. Objective Results

The segmental weighted signal-to-noise ratio (SNR) of the phase quantizer is illustrated in Fig. 6. The segmental SNR was calculated by averaging the SNR of the extracted waveforms. For each waveform, the SNR was computed using the quantized phase and nonquantized magnitude. The proposed system achieves approximately 14 dB SNR for as few as six bits for nonfiltered speech, and nearly 10 dB for modified intermediate reference system (M-IRS) [35] filtered speech.

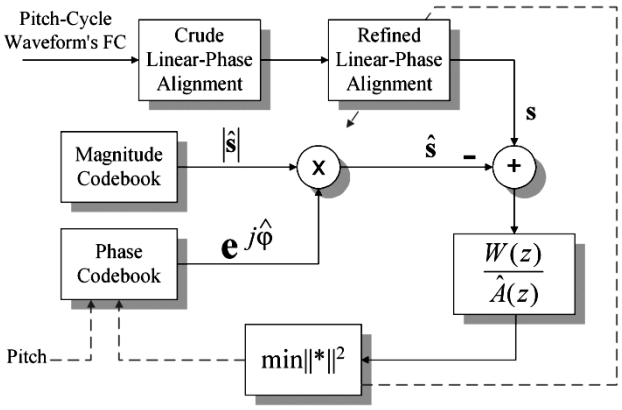


Fig. 5. Block diagram of the AbS dispersion phase vector quantization.

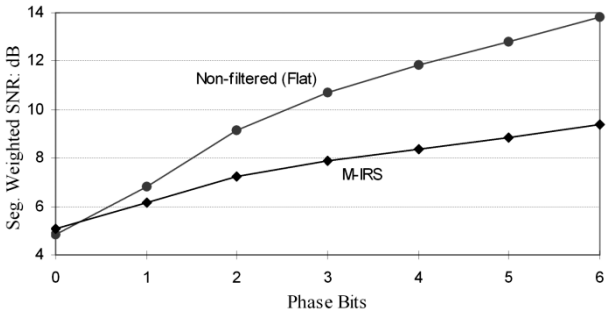


Fig. 6. Segmental weighted SNR of the phase VQ versus the number of bits, for M-IRS and for nonfiltered (flat) speech.

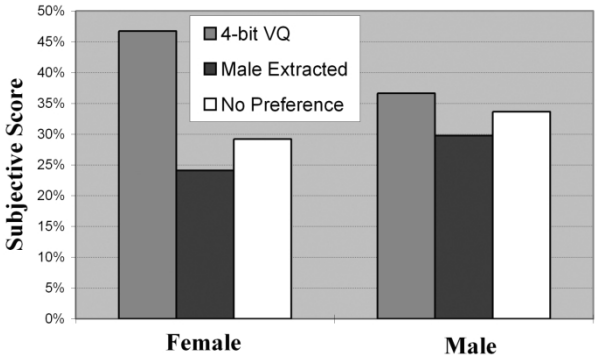


Fig. 7. Results of subjective A/B test for comparison between the four-bit phase VQ, and male extracted fixed phase.

D. Subjective Results

Recent WI coders have used a fixed dispersion phase extracted from male speakers [14], [19]. We have conducted a subjective A/B test to compare our dispersion phase VQ, using only four bits, to a male-extracted dispersion phase. The test data included 16 M-IRS speech sentences, eight of which are of female speakers, and eight 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 our WI system in which only the dispersion phase was quantized every 20 ms. Twenty one listeners participated in the test. The test results, illustrated in Fig. 7, show improvement in speech quality by quantizing the phase with a four-bit VQ. The improvement is larger for female speakers than for male. This may be due to the fact that for fe-

male speech there is a larger number of bits per vector sample, resulting in better waveform matching which is more perceivable particularly during transitions.

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

V. Parametric REW Quantization

Efficient REW quantization can benefit from two observations [25]: 1) the REW magnitude is typically an increasing function of frequency, which suggests that an efficient parametric representation may be used and 2) 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.

A. 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 polynomials [18]–[20]. 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 $\psi_i(\omega)$

$$R(\omega) = \sum_{i=0}^{I-1} \gamma_i \psi_i(\omega), \quad 0 \le \omega \le \pi \tag{21}$$

where ω is the angular frequency, and I is the representation order. The REW magnitude is typically an increasing function of 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. Consequently we model the REW by the following parametric representation, $\widehat{R}\ (\omega, \xi)$

$$\widehat{R}\,(\omega, \xi) = \sum_{i=0}^{I-1} \widehat{\gamma}_i\ (\xi)\psi_i(\omega), \quad 0 \le \omega \le \pi; \quad 0 \le \xi \le 1 \tag{22}$$

where $\widehat{\boldsymbol{\gamma}}\ (\xi) = [\widehat{\gamma}_0\ (\xi), \ldots, \widehat{\gamma}_{I-1}\ (\xi)]^T$ 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. Thus, $\widehat{R}\ (\omega, \xi)$ defines a 2-D surface whose cross sections for each value of ξ give a particular REW magnitude spectrum, which is defined merely by specifying a scalar parameter value.

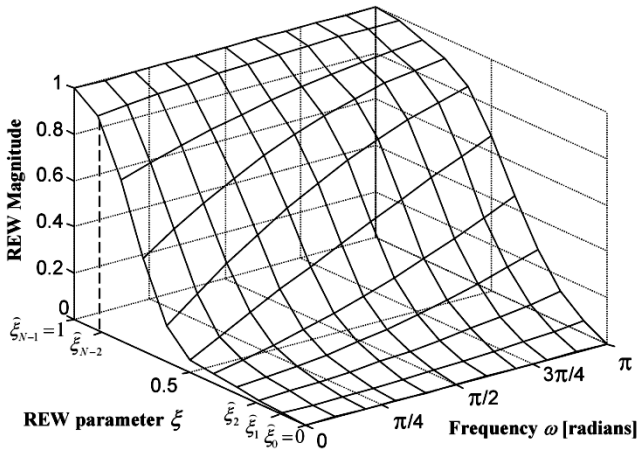


Fig. 8. REW parametric representation $\widehat{R}\ (\omega, \xi)$.

B. Piecewise Linear REW Representation

In order to have a simple representation that is computationally efficient and avoids excessive memory requirements, we model the 2-D surface by a piecewise linear parametric representation. Therefore, we introduce a set of N uniformly spaced spectra, $\{\widehat{R}\ (\omega, \widehat{\xi}_n)\}_{n=0}^{N-1}$, as shown in Fig. 8. (Such a set of functions is similar to the hand-tuned REW codebook in [19] and [20].) Then the parametric surface is defined by linear interpolation according to

$$\widehat{R}\,(\omega, \xi) = (1-\alpha)\ \widehat{R}\,(\omega, \widehat{\xi}_{n-1}) + \alpha\ \widehat{R}\,(\omega, \widehat{\xi}_n);$$
$$\widehat{\xi}_{n-1} \le \xi \le \widehat{\xi}_n;\ \alpha = \frac{\xi - \widehat{\xi}_{n-1}}{\Delta\,\widehat{\xi}};\ \Delta\,\widehat{\xi} = \widehat{\xi}_n - \widehat{\xi}_{n-1}\,. \tag{23}$$

Because this representation is linear, the coefficients of $\widehat{R}\ (\omega, \xi)$ are linear combinations of the coefficients of $\widehat{R}\ (\omega, \widehat{\xi}_{n-1})$ and $\widehat{R}\ (\omega, \widehat{\xi}_n)$. Hence,

$$\widehat{\boldsymbol{\gamma}}\ (\xi) = (1-\alpha)\ \widehat{\boldsymbol{\gamma}}_{n-1} + \alpha\ \widehat{\boldsymbol{\gamma}}_n \tag{24}$$

where $\widehat{\boldsymbol{\gamma}}_n$ is the coefficient vector of the nth REW magnitude representation

$$\widehat{\boldsymbol{\gamma}}_n = \widehat{\boldsymbol{\gamma}}\ (\widehat{\xi}_n). \tag{25}$$

C. REW Modeling

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

$$D(R, \widehat{R}\ (\xi)) = \int_0^{\pi} |R(\omega) - (1-\alpha)\ \widehat{R}\,(\omega, \widehat{\xi}_{n-1}) - \alpha\ \widehat{R}\,(\omega, \widehat{\xi}_n)|^2\, d\omega. \tag{26}$$

From orthonormality, the distortion is equal to

$$D(R, \widehat{R}\ (\xi)) = \|\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}\ (\xi)\|^2$$

$$= \|\boldsymbol{\gamma} - (1-\alpha)\,\widehat{\boldsymbol{\gamma}}_{n-1} - \alpha\,\widehat{\boldsymbol{\gamma}}_{n}\,\|^2. \quad (27)$$

The optimal interpolation factor that minimizes the MSE is

$$\alpha_{\text{opt}} = \frac{(\widehat{\boldsymbol{\gamma}}_n - \widehat{\boldsymbol{\gamma}}_{n-1})^T(\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}_{n-1})}{\|\,\widehat{\boldsymbol{\gamma}}_n - \widehat{\boldsymbol{\gamma}}_{n-1}\,\|^2} \quad (28)$$

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

$$\xi(\boldsymbol{\gamma}) = (1-\alpha_{\text{opt}})\,\widehat{\xi}_{n-1} + \alpha_{\text{opt}}\,\widehat{\xi}_{n}\,. \quad (29)$$

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.

2) Weighted Distortion: Commonly in speech coding, the magnitude is quantized using a weighted distortion measure. In this case, the weighted distortion between the input and the parametric representation modeled spectra is equal to

$$D_w(R, \widehat{R}\,(\xi)) = \int_0^{\pi} |R(\omega) - \widehat{R}\,(\omega,\xi)|^2 W(\omega)\,d\omega$$
$$= (\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}\,(\xi))^T \boldsymbol{\Psi}(W(\omega))(\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}\,(\xi)) \quad (30)$$

where $\boldsymbol{\Psi}(W(\omega))$ is the weighted correlation matrix of the orthonormal functions, its elements are

$$\Psi_{i,j}(W(\omega)) = \int_0^{\pi} W(\omega)\psi_i(\omega)\psi_j(\omega)\,d\omega \quad (31)$$

where $\boldsymbol{\gamma}$ is the input coefficient vector and $\widehat{\boldsymbol{\gamma}}\,(\xi)$ is the modeled parametric coefficient vector. The optimal parameter that minimizes (30) is given by

$$\alpha_{\text{opt}} = \frac{(\widehat{\boldsymbol{\gamma}}_n - \widehat{\boldsymbol{\gamma}}_{n-1})^T \boldsymbol{\Psi} (\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}_{n-1})}{(\widehat{\boldsymbol{\gamma}}_n - \widehat{\boldsymbol{\gamma}}_{n-1})^T \boldsymbol{\Psi} (\widehat{\boldsymbol{\gamma}}_n - \widehat{\boldsymbol{\gamma}}_{n-1})} \quad (32)$$

and the respective optimal parameter value is computed using (29). Alternatively, in order to eliminate using the matrix $\boldsymbol{\Psi}$, and to benefit from the orthonormal function simplification, given in (27), the scalar product may be redefined to incorporate the time-varying spectral weighting. The respective orthonormal basis functions then satisfy

$$\int_0^{\pi} W(\omega)\psi_i(\omega)\psi_j(\omega)\,d\omega = \delta(i-j) \quad (33)$$

where $\delta(i-j)$ denotes the Kroneker delta. The respective parameter vector is given by

$$\gamma = \int_0^{\pi} W(\omega) R(\omega) \boldsymbol{\psi}(\boldsymbol{\omega})\,d\omega \quad (34)$$

where $\boldsymbol{\psi}(\boldsymbol{\omega}) = [\psi_0, \psi_1, \ldots, \psi_{I-1}]^T$ is an Ith dimensional vector of time-varying orthonormal functions.

D. REW Quantization

1) Full Complexity Spectral Quantization Scheme: A novel AbS REW parameter VQ paradigm is illustrated in Fig. 9. An excitation vector $\hat{c}_{ij}(m)$ $(m = 1, \ldots, M)$ is selected from the VQ codebook and is fed through a synthesis filter to obtain a parameter vector $\hat{\xi}(m)$ (synthesized quantized) which is

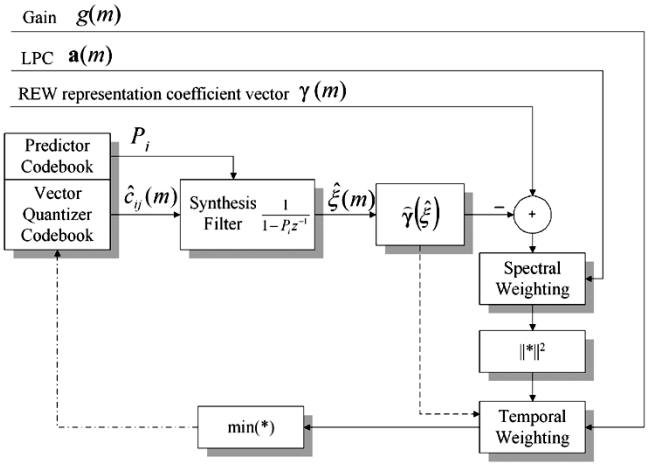


Fig. 9. REW parametric representation AbS VQ.

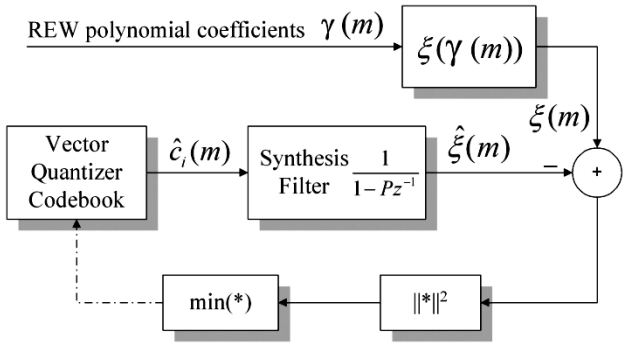


Fig. 10. REW parametric representation AbS VQ.

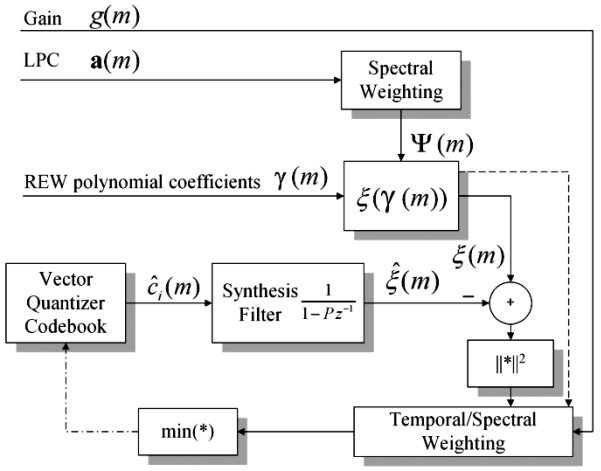


Fig. 11. REW parametric representation simplified weighted AbS VQ.

then mapped to quantized a representation coefficient vectors $\widehat{\boldsymbol{\gamma}}\,(\hat{\xi}(m))$. This is compared with a sequence of input representation coefficient vectors $\boldsymbol{\gamma}(m)$ and each is spectrally weighted. Each spectrally weighted error is then temporally weighted, and a distortion measure is obtained. A search through all candidate excitation vectors determines an optimal choice. The synthesis filter in Fig. 9 can be viewed as a first order predictor in a feedback loop. By allowing the value of the predictor parameter P to change, it becomes a "switched-predictor" scheme. 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. Two codebooks are used, one corresponding to each of two predictor coefficients, P_1 and P_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 $\boldsymbol{\gamma}(m)$. The search for the minimal weighted mean squared error (WMSE) is performed over all the vectors, $\hat{c}_{ij}(m)$, of the two codebooks for $i = 1, 2$. The quantized REW function coefficients vector, $\widehat{\boldsymbol{\gamma}}\ (\hat{\xi}(m))$, is a function of the quantized parameter $\widehat{\xi}\ (m)$, which is obtained by passing the quantized vector, $\hat{c}_{ij}(m)$, through the synthesis filter using the coefficient P_i for $i = 1$, or 2. 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.

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 a high rate, and a smooth representation surface $\widehat{R}\ (\omega, \xi)$, the total distortion is equal to the sum of a modeling distortion and a quantization distortion

$$\sum_{m=1}^{M} D_w(R(m), \hat{R}(\hat{\xi}(m))) = \sum_{m=1}^{M} D_w(R(m), \widehat{R}\ (\xi(m))) + \sum_{m=1}^{M} D_w(\widehat{R}\ (\xi(m)), \hat{R}(\hat{\xi}(m))). \tag{35}$$

The quantization distortion is related to the quantized parameter by

$$\sum_{m=1}^{M} D_w(\widehat{R}\ (\xi(m)), \hat{R}(\hat{\xi}(m))) = \sum_{m=1}^{M} (\widehat{\boldsymbol{\gamma}}\ (\xi(m)) - \hat{\boldsymbol{\gamma}}(\hat{\xi}(m)))^T \times \boldsymbol{\Psi}(W(m))(\widehat{\boldsymbol{\gamma}}\ (\xi(m)) - \hat{\boldsymbol{\gamma}}(\hat{\xi}(m))) \tag{36}$$

which, for the piecewise linear representation case, is equal to

$$\sum_{m=1}^{M} D_w(\widehat{R}\ (\xi(m)), \hat{R}(\hat{\xi}(m))) = \frac{1}{\Delta\ \widehat{\xi}^{\,2}} \sum_{m=1}^{M} \Delta\ \widehat{\boldsymbol{\gamma}}_n\ (\xi(m))^T \times \boldsymbol{\Psi}(W(m)) \Delta\ \widehat{\boldsymbol{\gamma}}_n\ (\xi(m))(\xi(m) - \hat{\xi}(m))^2 \tag{37}$$

where

$$\Delta\ \widehat{\boldsymbol{\gamma}}_n\ (\xi(m)) = \widehat{\boldsymbol{\gamma}}_n\ (\xi(m)) - \widehat{\boldsymbol{\gamma}}_{n-1}\ (\xi(m)). \tag{38}$$

The quantization distortion is linearly related to the REW parameter squared quantization error, $(\xi(m) - \hat{\xi}(m))^2$, and therefore justifies direct VQ of the REW parameter.

a) Simplified scheme, nonweighted distortion: The encoder maps the REW magnitude to an unvoicing parameter, and then quantizes the parameter by AbS VQ, as illustrated in Fig. 10. Initially, the magnitudes of the M REWs in the frame are mapped to coefficient vectors, $\{\boldsymbol{\gamma}(m)\}_{m=1}^{M}$. Then, for each coefficient vector, a search is performed to find the optimal representation parameter, $\xi(\boldsymbol{\gamma})$, using (29), to form an M-dimensional parameter vector for the current frame, $\{\xi(\boldsymbol{\gamma}(m))\}_{m=1}^{M}$. Finally, the parameter vector is encoded by AbS VQ. The decoded spectra, $\{\widehat{R}\,(\omega, \hat{\xi}(m))\}_{m=1}^{M}$, are obtained from the quantized parameter vector, $\{\hat{\xi}(m)\}_{m=1}^{M}$, using (23). 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.

b) Simplified scheme, weighted distortion: We may improve the quantization scheme to incorporate spectral and temporal weightings, as illustrated in Fig. 11. The REW parameter vector is first mapped to a REW parameter by minimizing a distortion, which is weighted by the coefficient spectral weighting matrix $\boldsymbol{\Psi}$, as described above. Then, the resulting REW parameter is used to compute a weighting, $w_s(\xi(m))$, which we choose to be the spectral sensitivity to the REW parameter squared quantization error, $(\xi(m) - \hat{\xi}(m))^2$, given by

$$w_s(\xi(m)) = \left(\frac{\partial\ \widehat{\boldsymbol{\gamma}}}{\partial \xi}\right)^T \boldsymbol{\Psi} \left(\frac{\partial\ \widehat{\boldsymbol{\gamma}}}{\partial \xi}\right)_{|_{\xi(m)}}. \tag{39}$$

For the piecewise linear representation case it is equal to

$$w_s(\xi(m)) = \frac{1}{\Delta\ \widehat{\xi}^{\,2}} \Delta\ \widehat{\boldsymbol{\gamma}}_n\ (\xi(m))^T \boldsymbol{\Psi}(W(m)) \Delta\ \widehat{\boldsymbol{\gamma}}_n\ (\xi(m)). \tag{40}$$

This derivative can be computed off line. Additionally, for the temporal weighting, a monotonic nonlinear function of the gain, denoted by $w_t(g(m))$, is used to give relatively large weight to waveforms with larger gain values. The AbS REW parameter quantization is computed by minimizing the combined spectrally and temporally weighted distortion

$$D\left(\{\xi(m)\}_{m=1}^{M}, \{\hat{\xi}(m)\}_{m=1}^{M}\right) = \sum_{m=1}^{M} w_t(g(m)) w_s(\xi(m)) \times (\xi(m) - \hat{\xi}(m))^2. \tag{41}$$

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.

VI. Dual Predictive SEW Optimization

Fig. 12 illustrates a dual predictive SEW AbS VQ scheme which uses two observables: 1) the quantized REW and 2) the past quantized SEW, to jointly predict the current SEW. Although we refer to the operator on each observable as a "predictor," in fact both are components of a single optimized estimator. The SEW and the REW are complex random vectors, and their sum is a residual vector having elements whose magnitudes have a mean value of unity. In low bit-rate WI coding, the relation between the SEW and the REW magnitudes was approximated by computing the magnitude of one as the unity complement of the other [14], [17]–[20]. Suppose $|\hat{\mathbf{r}}_M|$ denotes the spectral magnitude vector of the last quantized REW in the current frame. An "implied" SEW vector is calculated by

$$|\hat{\mathbf{s}}_{M,\text{implied}}| = 1 - |\hat{\mathbf{r}}_M| \tag{42}$$

and from which the mean vector is to be removed. Vectors whose means are removed are denoted with an apostrophe. Then, we compute a (mean-removed) estimated "implied" SEW magnitude vector, $|\tilde{\mathbf{s}}'_{M,\text{implied}}|$, using a diagonal estimation matrix $\mathbf{P}_{\text{REW}}$,

$$|\tilde{\mathbf{s}}'_{M,\text{implied}}| = \mathbf{P}_{\text{REW}}|\hat{\mathbf{s}}'_{M,\text{implied}}|. \tag{43}$$

Additionally, a "self-predicted" SEW vector is computed by multiplying the delayed quantized SEW vector, $|\hat{\mathbf{s}}'_0|$, by a diagonal prediction matrix $\mathbf{P}_{\text{SEW}}$. The predicted (mean-removed) SEW vector, $|\tilde{\mathbf{s}}'_M|$, is given by

$$|\tilde{\mathbf{s}}'_M| = \mathbf{P}_{\text{REW}}|\hat{\mathbf{s}}'_{M,\text{implied}}| + \mathbf{P}_{\text{SEW}}|\hat{\mathbf{s}}'_0|. \tag{44}$$

The quantized vector, $\hat{\mathbf{c}}_M$, is determined by an AbS search according to

$$\hat{\mathbf{c}}_M = \underset{\mathbf{c}_i}{\text{argmin}}\{(|\mathbf{s}'_M| - |\tilde{\mathbf{s}}'_M| - \mathbf{c}_i)^T\mathbf{W}_M \times (|\mathbf{s}'_M| - |\tilde{\mathbf{s}}'_M| - \mathbf{c}_i)\} \tag{45}$$

where $\mathbf{W}_M$ is the diagonal spectral weighting matrix. The (mean-removed) quantized SEW magnitude, $|\hat{\mathbf{s}}'_M|$, is the sum of the predicted SEW vector, $|\tilde{\mathbf{s}}'_M|$, and the codevector $\hat{\mathbf{c}}_M$

$$|\hat{\mathbf{s}}'_M| = |\tilde{\mathbf{s}}'_M| + \hat{\mathbf{c}}_M. \tag{46}$$

The EWI coder uses the optimized SEW, $\mathbf{s}_{M,\text{opt}}$, and the optimized weighting, $\mathbf{W}_{M,\text{opt}}$, for the AbS SEW quantization.

In order to exploit the information about the pitch, and the voicing level, we have partitioned the possible pitch range into six subintervals, and the REW parameter range into three, and generated 18 codebooks, one for each pair of pitch range and unvoicing level. The pitch and the unvoicing level determine which codebook is searched. Each codebook has associated with it two mean vectors, and two diagonal prediction matrices. To improve the coder robustness and the synthesis smoothness, the cluster used for the training of each codebook overlaps with those of the codebooks for neighboring ranges. Since each quantized target vector may have a different value of the removed mean, the quantized mean is added temporarily to the filter memory after the state update, and the next quantized vector's mean is subtracted from it before filtering is performed.

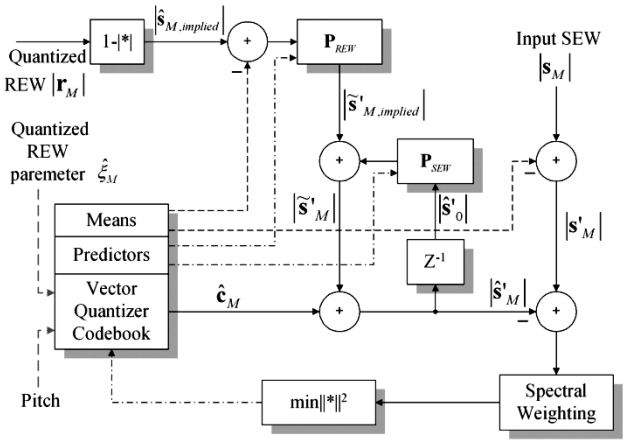


Fig. 12. Block diagram of the dual predictive AbS SEW VQ.

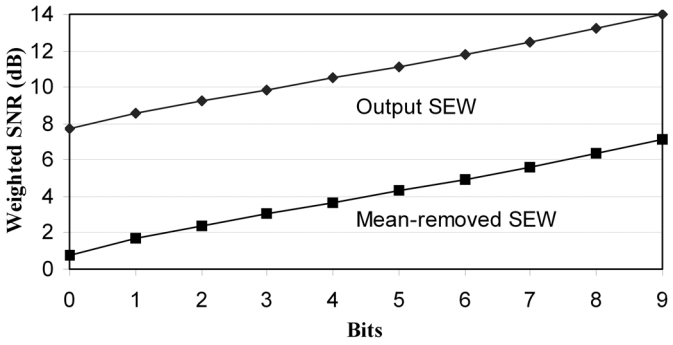


Fig. 13. Weighted SNR for dual predictive AbS SEW VQ.

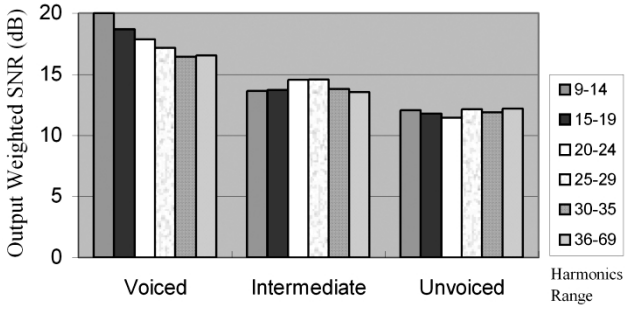


Fig. 14. Output weighted SNR for the 18 codebooks, nine-bit AbS SEW VQ.

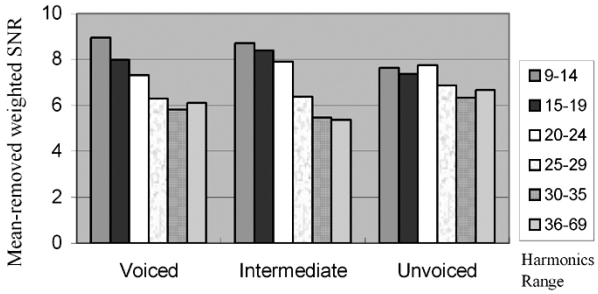


Fig. 15. Mean-removed SEWs Weighted SNR for the 18 codebooks, nine-bit AbS SEW VQ.

The output weighted SNR, and the mean-removed weighted SNR, of the scheme, for M-IRS [35] filtered speech, are illustrated in Fig. 13. Evidently, a very high SNR is achieved

with a relatively small number of bits. The weighted SNR of each codebook, for the nine-bit case, is illustrated in Fig. 14. The differences in SNR between the three REW parameter ranges is dominated by the different means. The respective mean-removed weighted SNR of each codebook is illustrated in Fig. 15. Within each voicing range, the differences in SNR for different pitch ranges are mainly due to the prediction gain and to the number of bits per vector sample, which decreases as the number of harmonics increases.

Example for the two predictors for three REW parameter ranges is illustrated in Fig. 16. For a voiced segment, the SEW predictor is dominant and the REW predictor is less important since its input variations in this range are very small. As the voicing decreases, the contribution of the SEW predictor decreases, and the REW predictor becomes the dominant contributor at the lower part of the spectrum. Both predictors give decreasing contributions as the voicing decreases from the intermediate range to the unvoiced range.

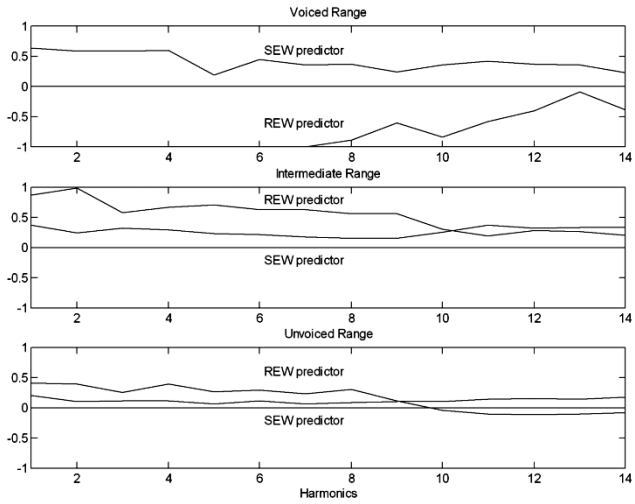


Fig. 16. Predictors for three REW parameter ranges.

VII. Pitch Search

The pitch search consists of a spectral domain search employed every 10 ms and a temporal domain search employed every 2 ms, as illustrated in Fig. 17. The spectral domain pitch search is based on harmonic matching [2], [3], [30]. 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 (see (47) at the bottom of the page) where Δ is some incremental segment used in the summations for computational simplicity, and $0 \leq N_j \leq \lfloor 160/\Delta \rfloor$. Then, every 10 ms a weighted-mean pitch value is calculated by

$$P_{\text{mean}} = \sum_{i=1}^{5} \rho(n_i) P(n_i) \Bigg/ \sum_{i=1}^{5} \rho(n_i) \tag{48}$$

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

VIII. 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, shown in Fig. 18. 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 nonlinear function of the temporal gain, $g(m)$, which gives greater emphasis to larger gain values. Two codebooks 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.

IX. Bit Allocation

The bit allocation for the 2.8 kb/s EWI coder is given in Table I. The frame length is 20 ms, and ten waveforms are extracted per frame. The *line spectral frequencies* (LSFs) are coded using predictive multi-stage VQ (MSVQ), having two stages of ten bits each, a two-bit increase compared to the previous version of our coder [22], [23]. This bit increase improves the resolution of the spectral envelope and therefore yields whiter residual which improves the modeling, and therefore improves speech quality, most notably in the transitions. The 10th dimensional log-gain vector is quantized using nine bit AbS VQ [22], [23], including one bit for the switch prediction.

$$\begin{aligned} P(n_i) &= \underset{\tau, N_1, N_2}{\arg\max} \{\rho(n_i, \tau, N_1, N_2)\} \\ &= \underset{\tau, N_1, N_2}{\arg\max} \left\{ \frac{\sum_{n=n_i-N_1\Delta}^{n_i+\tau+N_2\Delta} s_w(n) s_w(n-\tau)}{\sqrt{\sum_{n=n_i-N_1\Delta}^{n_i+\tau+N_2\Delta} s_w(n) s_w(n)} \sqrt{\sum_{n=n_i-N_1\Delta}^{n_i+\tau+N_2\Delta} s_w(n-\tau) s_w(n-\tau)}} \right\} \end{aligned} \tag{47}$$

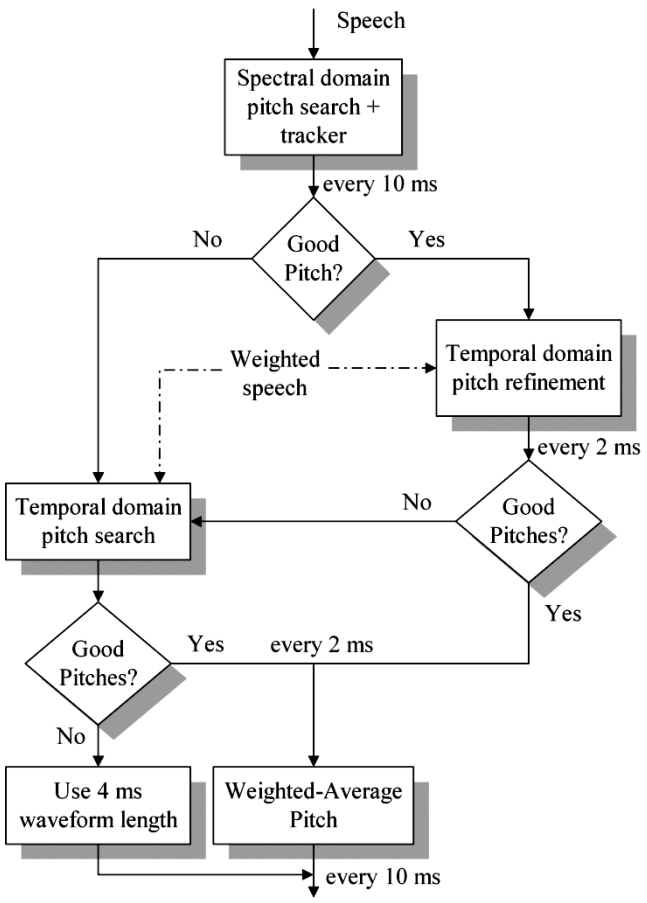


Fig. 17. Pitch search of the EWI coder.

The pitch is coded twice per frame. A fixed SEW phase was trained for each one of the eighteen pitch-voicing ranges [21], as explained in Section IV.

X. Subjective Test Results

We have conducted a subjective A/B test to compare our 2.8 kb/s EWI coder to the G.723.1. The test data included 24 M-IRS [35] 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 Tables II and III, indicate that the subjective quality of the 2.8 kb/s EWI exceeds that of G.723.1 at 5.3 kb/s, and it is slightly better than that of G.723.1 at 6.3 kb/s. The EWI preference is higher for male than for female speakers. In addition, we have done extensive listening tests with noisy speech and found that the EWI coder is robust to various noise conditions.

XI. Summary and Conclusions

We have found several new techniques that enhance the performance of the WI coder, and allow for better coding efficiency. The most significant of these, reported here, AbS optimization of the SEW, AbS vector-quantization of the dispersion-phase, dual-predictive AbS quantization of the SEW, efficient parameterization of the REW magnitude, AbS VQ of the REW parameter, a special pitch search for transitions, and switched-predictive AbS gain VQ. These features improve the algorithm and its robustness. Subjective test results indicate that the performance of the 2.8 kb/s EWI coder slightly exceeds that of G.723.1 at 6.3 kb/s and therefore EWI achieves very close to toll quality, at least under clean speech conditions.

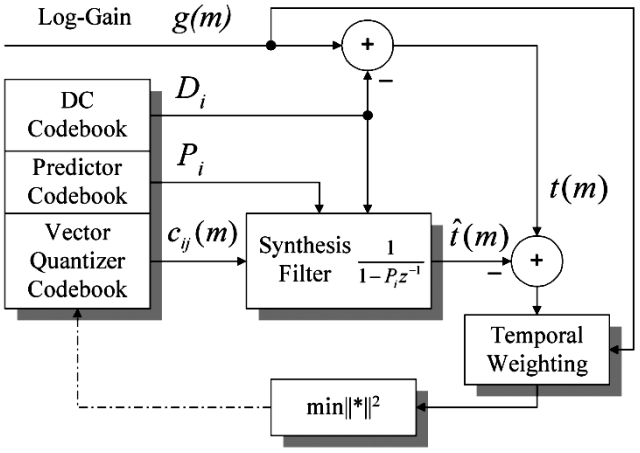


Fig. 18. Switched-predictive analysis-by-synthesis gain VQ using temporal weighting.

TABLE I
2.8 kb/s EWI Bit Allocation

Parameter	Bits / Frame	Bits / second
LPC	10+10 = 20	1000
Pitch	2x6 = 12	600
Gain	9	450
SEW magnitude	8	400
REW magnitude	7	350
Total	**56**	**2800**

TABLE II
A/B Test: EWI Versus 5.3 kb/s G.723.1

Test	*2.8 kb/s WI*	*5.3 kb/s G.723.1*	*No Preference*
Female	*40.3%*	*33.3%*	*26.4%*
Male	*48.6%*	*24.3%*	*27.1%*
Total	***44.4%***	***28.8%***	***26.7%***

TABLE III
A/B Test: EWI Versus 6.3 kb/s G.723.1

Test	*2.8 kb/s WI*	*6.3 kb/s G.723.1*	*No Preference*
Female	*38.2%*	*36.8%*	*25.0%*
Male	*43.1%*	*31.9%*	*25.0%*
Total	**40.6%**	**34.4%**	**25.0%**

Acknowledgment

The authors would like to thank W. Bastiaan Kleijn for the insightful comments during the course of this research and Kenneth Rose for the interesting discussions during the phase vector quantization design.

References

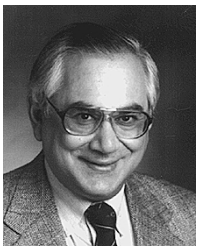
[1] B. S. Atal and M. R. Schroeder, "Stochastic coding of speech at very low bit-rate," in *Proc. Int. Conf. Comm.*, Amsterdam, The Netherlands, 1984, pp. 1610–1613.

[2] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam, The Netherlands: Elsevier, 1995, ch. 4, pp. 121–173.
[3] D. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1223–1235, Aug. 1988.
[4] A. V. McCree and T. P. Barnwell III, "A new mixed excitation LPC vocoder," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 593–596, 1991.
[5] ——, "A mixed excitation lpc vocoder model for low bit-rate speech coding," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 242–250, July 1995.
[6] C. Laflamme, R. Salami, R. Matmti, and J.-P. Adoul, "Harmonic-stochastic excitation (HSX) speech coding below 4 kbit/s," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1996, pp. 204–207.
[7] Y. Shoham and A. Gersho, "Pitch synchronous transform coding of speech at 9.6 kb/s based on vector quantization," in *Proc. Int. Conf. Commun.*, Amsterdam, 1984, pp. 1179–1182.
[8] W. B. Kleijn, "Continuous representations in linear predictive coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1991, pp. 201–203.
[9] Y. Shoham, "High quality speech coding at 2.4 to 4.0 kb/s based on time-frequency-interpolation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 167–170, 1993.
[10] I. S. Burnett and R. J. Holbeche, "A mixed prototype waveform/celp coder for sub 3 kb/s," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, pp. 175–178, 1993.
[11] W. B. Kleijn, "Encoding speech using prototype waveforms," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 386–399, Oct. 1993.
[12] W. B. Kleijn and J. Haagen, "Transformation and decomposition of the speech signal for coding," *IEEE Signal Processing Lett.*, vol. 1, no. 9, pp. 136–138, 1994.
[13] ——, "A speech coder based on decomposition of characteristic waveforms," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 508–511, 1995.
[14] ——, "Waveform interpolation for coding and synthesis," in *Speech Coding Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds: Elsevier, 1995, ch. 5, pp. 175–207.
[15] I. S. Burnett and G. J. Bradley, "New techniques for multi-prototype waveform coding at 2.84 kb/s," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1995, pp. 261–263.
[16] ——, "Low complexity decomposition and coding of prototype waveforms," in *IEEE Workshop Speech Coding Telecommunications*, 1995, pp. 23–24.
[17] I. S. Burnett and D. H. Pham, "Multi-prototype waveform coding using frame-by-frame analysis-by-synthesis," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 1567–1570, 1997.
[18] W. B. Kleijn, Y. Shoham, D. Sen, and R. Haagen, "A low-complexity waveform interpolation coder," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 212–215, 1996.
[19] Y. Shoham, "Very low complexity interpolative speech coding at 1.2 to 2.4 kb/s," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 1599–1602, 1997.
[20] ——, "Low-complexity speech coding at 1.2 to 2.4 kb/s based on waveform interpolation," *Int. J. Speech Technol.*, pp. 329–341, May 1999.
[21] O. Gottesman, "Dispersion phase vector quantization for enhancement of waveform interpolative coder," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, 1999, pp. 269–272.
[22] O. Gottesman and A. Gersho, "Enhanced waveform interpolative coding at 4 kb/s," in *IEEE Speech Coding Workshop*, Finland, 1999, pp. 90–92.
[23] ——, "Enhanced analysis-by-synthesis waveform interpolative coding at 4 kb/s," in *EUROSPEECH'99*, Hungary, 1999, pp. 1443–1446.
[24] ——, "High quality enhanced waveform interpolative coding at 2.8 kb/s," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Istanbul, Turkey, June 2000, pp. 1363–1366.
[25] ——, "Enhancing waveform interpolative coding with weighted REW parametric quantization," in *IEEE Workshop on Speech Coding Proc.*, Sept. 2000, pp. 50–52.
[26] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. London, U.K.: Academic, 1989.
[27] G. Kubin, B. S. Atal, and W. B. Kleijn, "Performance of noise excitation for unvoiced speech," in *Proc. IEEE Workshop Speech Coding Telecommun.*, Sainte-Adele, QC, Canada, 1993, pp. 35–36.
[28] J. H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 59–71, Jan. 1995.
[29] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer, 1992.
[30] E. Shlomot, V. Cuperman, and A. Gersho, "Hybrid coding of speech at 4 kb/s," in *IEEE Speech Coding Workshop*, 1997, pp. 37–38.
[31] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 49, no. 2, pp. 583–590, 1971.
[32] Y. Jiang and V. Cuperman, "Encoding prototype waveforms using a phase codebook," in *IEEE Workshop Speech Coding Telecommunications*, 1995, pp. 21–22.
[33] W. R. Gardner and B. D. Rao, "Noncausal all-pole modeling of voiced speech," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 1–10, Jan. 1997.
[34] X. Sun *et al.*, "Phase modeling of speech excitation for low bit-rate sinusoidal transform coding," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 1691–1694, 1997.
[35] *ITU Recommend. P.830, Subjective Performance Assessment Telephone Band Wideband Digital Codecs*, Feb. 1996.

Oded Gottesman (S'88–M'01) received the B.Sc. (cum laude) degree in electrical engineering from Ben Gurion University, Israel, in 1988, the M.Sc. degree in electrical engineering from Drexel University, Philadelphia, PA, in 1993, and the Ph.D. degree in electrical engineering from the University of California at Santa Barbara (UCSB), in 2000.

From October 1989 to September 1990, he was with Efrat Future Technologies, Israel, where he carried out research and development in the area of speech compression and digital signal processing for telephony. From January to September 1992 he worked as Consultant with AT&T Bell Labs, Murray Hill, NJ, where he performed his M.Sc. research in the area of low delay wideband speech coding. From April to December 1993 he was the Audio Group Manager with Optibase Inc., Israel, where he carried out research and development in the area of speech compression and digital signal processing for teleconferencing. From January 1994 to September 1995, he was the leader of the speech coding group in DSP Communications, Israel, where he carried out research and development of speech coding and speech enhancement algorithms for cellular phone and other applications. He is currently a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, UCSB. His research interests are in source and channel coding, speech, audio and video coding and processing. He has pending patents on speech coding.

Dr. Gottesman was co-recipient of the Ericsson-Nokia Best Paper Award at the IEEE Workshop on Speech Coding, Finland, 1999. He received an award for academic distinction from the Israeli Parliament in 1988.



Allen Gersho (S'58–M'64–SM'78–F'81) received the B.S. degree from the Massachusetts Institute of Technology, Cambridge, in 1960, and the Ph.D. degree from Cornell University, Ithaca, NY, in 1963.

He was with Bell Laboratories from 1963 to 1980. He is currently a Research Professor of electrical and computer engineering at the University of California at Santa Barbara. His current research activities are in signal compression methodologies and algorithm development for speech, audio, image, and video coding. He holds patents on speech coding, quantization, adaptive equalization, digital filtering, and modulation and coding for voiceband data modems. He is co-author (with R. M. Gray) of *Vector Quantization and Signal Compression* (Norwell, MA: Kluwer, 1992) and co-editor of two books on speech coding.

Dr. Gersho served as a member of the Board of Governors of the IEEE Communications Society from 1982 to 1985 and as a member of various IEEE technical, award, and conference management committees. He has served as Editor of IEEE COMMUNICATIONS MAGAZINE and Associate Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS. He received NASA Tech Brief Awards for technical innovation in 1987, 1988, and 1992. In 1980, he was co-recipient of the Guillemin-Cauer Prize Paper Award from the Circuits and Systems Society. He received the Donald McClennan Meritorious Service Award from the IEEE Communications Society in 1983, and in 1984, he was awarded an IEEE Centennial Medal. In 1992, he was co-recipient of the 1992 Video Technology Transactions Best Paper Award from the IEEE Circuits and Systems Society. In 1999, he was co-recipient of the Ericsson-Nokia Best Paper Award in the IEEE Workshop on Speech Coding. He was awarded an IEEE Third Millennium Medal in 2000.