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 compo-

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,

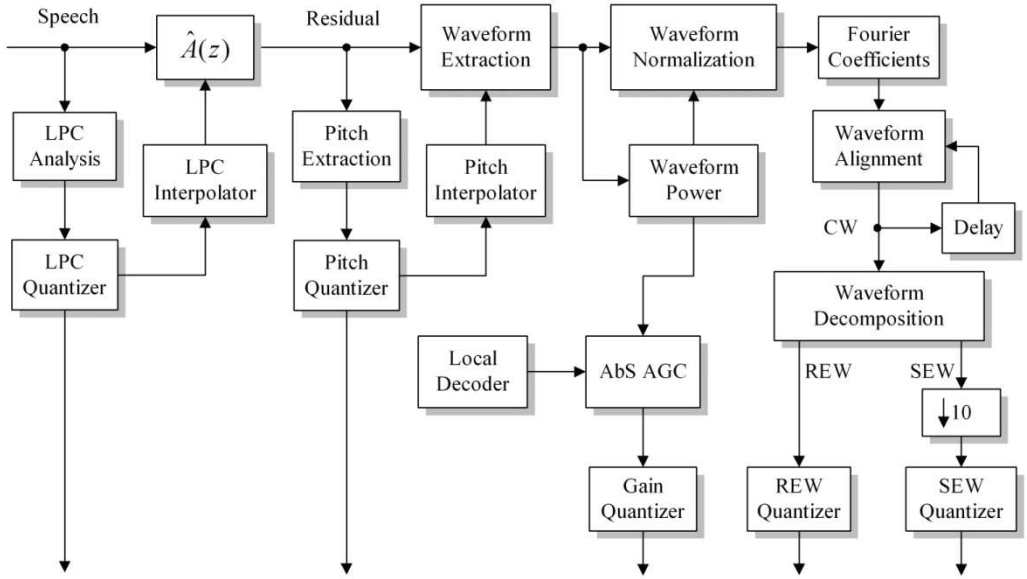

Fig. 1. Block diagram of the EWI encoder.

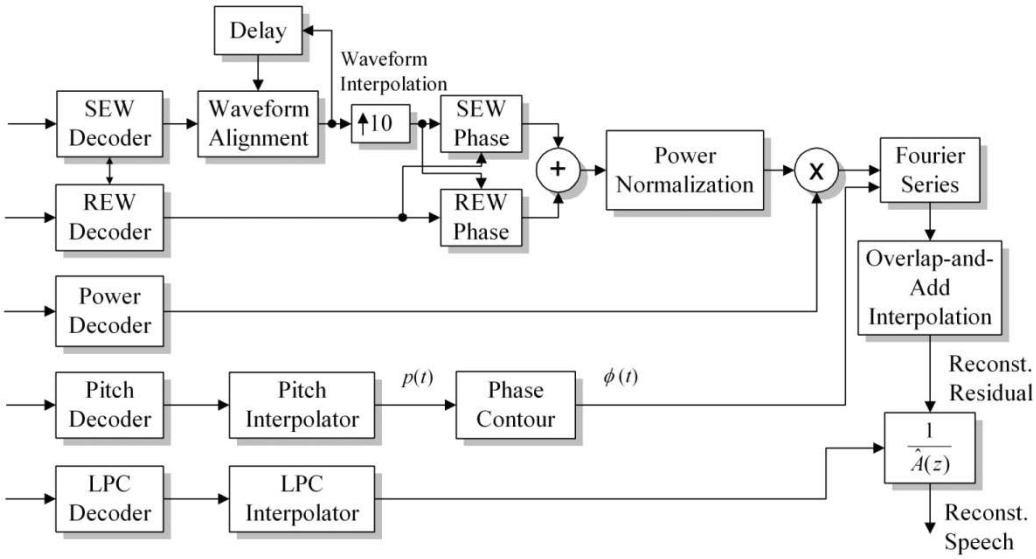

Fig. 2. Block diagram of the EWI decoder.

(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 ade-malized, 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

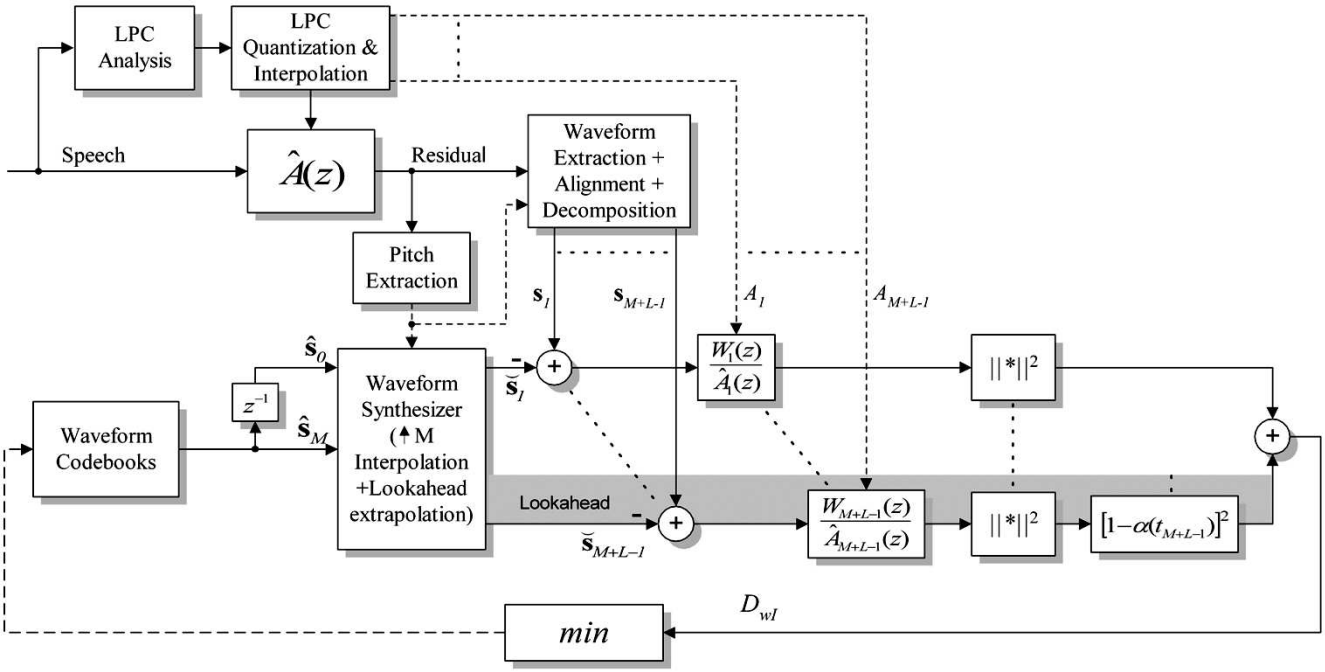

Fig. 3. Block diagram of the AbS SEW vector quantization.

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 non-periodic 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

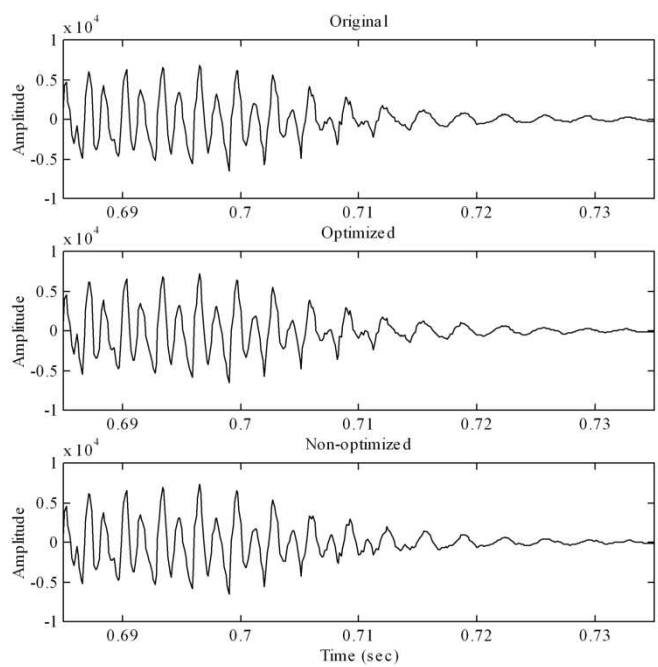

Fig. 4. Example for the improved interpolation by SEW optimization during nonstationary speech segment.

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 *accumu-*

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

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

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

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;
- $|\hat{\mathbf{s}}|$ 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

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.