A Multiband Excited Waveform-Interpolated 2.35-kbps Speech Codec for Bandlimited Channels

F. C. A. Brooks, *Member, IEEE,* and Lajos Hanzo, *Senior Member, IEEE*

***Abstract*—Following a brief portrayal of the activities in 2.4-kbps speech coding, a wavelet-based pitch detector is invoked, which reduces the complexity of conventional autocorrelation-based pitch detectors, while ensuring smooth pitch trajectory evolution. This scheme is incorporated in a waveform-interpolated codec, which uses voiced–unvoiced (V/U) classification, and instead of simple Dirac pulses, an unconventional zinc basis function excitation is employed for modeling the voiced excitation. The required zinc-function parameters are determined in an analysis-by-synthesis loop, and for the sake of smooth waveform evolution and reduced complexity, a focused search strategy and a few further suboptimum restrictions are imposed without seriously affecting the speech quality. This baseline codec operates at a rate of 1.9 kbps, but it suffers from slight buzziness during the periods of excessive voicing. This impediment is then mitigated by invoking a mixed V/U multiband excitation, which slightly increases the bit rate to 2.35 kbps due to the transmission of the 3-b voicing strength code in each of the three excitation bands.**

***Index Terms*—Low-rate speech coding, multiband excitation, waveform-interpolated speech coding, wavelet-based pitch estimation.**

I. Overview And Background

THE STATE-OF-ART of speech compression was documented in a range of excellent monographs by O'Shaughnessy [1], Furui [2], Kondoz [3], and Kleijn and Haagen [4] as well as in a tutorial review by Gersho [5]. More recently, the 5.6-kb/s half-rate GSM quadruple-mode vector sum excited linear predictive (VSELP) speech codec standard developed by Gerson *et al.* [6] was approved, while in Japan the 3.45-kb/s half-rate JDC speech codec invented by Ohya *et al.* [7] using the pitch synchronous innovation (PSI) CELP principle was standardized. Other recently investigated schemes are prototype waveform interpolation (PWI) proposed by Kleijn [8], multiband excitation (MBE) suggested by Griffin *et al.* [9], and interpolated zinc function prototype excitation (IZFPE) codecs advocated by Hiotakakos and Xydeas [10]. Last, the standardization of the 2.4-kbps DoD codec led to intensive research in this very low-rate range.

The seven 2.4-kbps DoD candidate coders can be divided into several categories. These categories were multiband excitation (MBE) [9], [11] and sinusoidal coders [12], [13], with the remaining candidate coders highlighted in [4], [14], and [15]. Following a set of rigorous comparative tests, the mixed excitation linear predictive (MELP) codec by Texas Instruments was selected for standardization [16].

Against this background, our elaborations in this treatise are centered around the well-established low bit rate speech compression technique of waveform interpolation (WI), pioneered by Kleijn [8]. In waveform interpolation, a characteristic waveform, which is also referred to as the prototype waveform, is periodically located in the original speech signal. Between these selected prototype segments, smoothly evolving interpolation is employed in order to reproduce the continuous synthesized speech signal. The interpolation can be performed in either the frequency or time domain, distinguishing two basic interpolation subclasses. Since only the prototype segments have to be encoded, the required bit rate is low, while maintaining good perceptual speech quality.

Most WI architectures rely on Kleijn's frequency-domain approach [4], [8], although there are schemes, such as that proposed by Hiotakakos and Xydeas [10] as well as Hiwasaki and Mano [17], which employ time-domain coding. A complication with any WI scheme is the need for interpolation between two prototype segments, which have different lengths. This paper uses a parametric excitation, which permits simple time-domain interpolation, as it will become explicit during our further discourse.

In traditional vocoders the decision as regards to the extent and nature of voicing in a speech segment is critical. Pitch estimation is an arduous task due to a number of factors, such as the nonstationary nature of speech, the effect of the vocal tract on the pitch frequency, and the presence of noise. Incorrect voicing decisions cause distortion in the reconstructed speech, and distortion is also apparent if the common phenomenon of pitch doubling occurs. Thus, for any low bit rate speech codec, the pitch detector chosen is vital in determining the resulting synthesized speech quality.

Recently, the wavelet transform has been applied to the task of pitch estimation [18], [19]. The wavelet approach to pitch estimation is event-based, implying that both the pitch period and the instant of glottal closure are determined. In the proposed speech codec, we employed a wavelet-based pitch detector.

Apart from the wavelet-based pitch estimation, this paper additionally investigates the low bit rate technique of multiband and mixed excitation, which is used in conjunction with the WI scheme. Multiband and mixed excitation both attempt to reduce the artifact termed "buzziness" by eliminating the binary classification into entirely voiced or unvoiced segments. This "buzzy" quality is particularly apparent in portions of speech,

Manuscript received November 17, 1998; revised April 7, 1999. This work was supported by the European Community, Brussels, Belgium, and the Engineering and Physical Sciences Research Council, Swindon, U.K.

The authors are with the Department of Electrical and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: lh@ecs.soton.ac.uk).

Publisher Item Identifier S 0018-9545(00)03689-6.

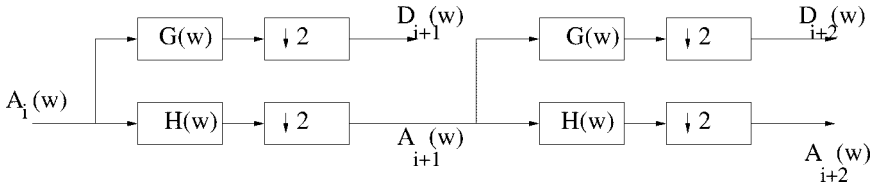

Fig. 1. Pyramidal algorithm for multiresolution analysis.

which have dominant voicing in some frequency regions, but dominant noise in other frequency bands of the spectrum.

We commence our discussions in Section II by a brief overview of wavelets, specifically the polynomial spline wavelets introduced by Mallat and Zhong [20]. The paper proceeds by applying the theory of wavelets to pitch estimation, which was proposed before for example in [18], [19], and [24]. However, we refined these wavelet-based pitch-estimation techniques and incorporated them in a correlation-based pitch detector, which allowed us to achieve a substantial pitch-estimation complexity reduction, as it will be demonstrated in Fig. 7. Subsequently, we introduced the wavelet-based pitch detector into a waveform interpolation speech codec. In Section III, we highlight the concept of the pitch prototype segment selection process of Fig. 8 and the features of the zinc basis function excitation of Fig. 6, which was originally proposed by Sukkar *et al.* [29] and dramatically refined by Hiotakakos and Xydeas [10]. Our contribution in the zinc-function excitation (ZFE) optimization is a further refinement of the technique, leading to the realization that the position of these ZFE pulses does not have to be explicitly signaled to the decoder, which reduces the bit rate. We then refine the processing of voiced–unvoiced (V/U) transitions and discuss the process of smooth excitation interpolation between prototype segments, as will be highlighted in Fig. 8 and characterize the zinc-excited codec performance. Finally, we proposed combining the well-known technique of mixed V/U multiband excitation [9] (MBE) with our ZFE-based codec in Fig. 10 of Section IV, in order to reduce the binary V/U decision induced "buzziness" and, hence, to further improve the reconstructed speech quality. We summarized our findings in Sections V and VI. Let us now focus our attention on the proposed wavelet-based pitch detector.

II. Wavelet-Based Pitch Estimation

In recent years, wavelets have stimulated significant research interests in a variety of applications. Historically, the theory of wavelets was recognized as a distinct discipline in the early 1980's. Daubechies [21] and Mallat [22] have substantially advanced this field in various signal processing applications. In this paper, wavelets are harnessed to reduce the computational complexity and improve the accuracy of an autocorrelation function (ACF)-based pitch detector. We commence with a brief introduction to wavelet theory.

A wavelet is an arbitrary function, which obeys certain conditions [21] that allow it to represent a signal $f(x)$ by a series of basis functions, which is described by

$$f(x) = \sum_{jk} d_{jk}\psi_{jk}(x) \tag{1}$$

where d_{jk} are the coefficients of the decomposition and ψ_{jk} are the basis functions for integers j and k.

The wavelet transform can be used to analyze a signal $f(x)$, but unlike the short-time Fourier transform, its localization varies over the time-frequency space. This flexibility in resolution makes it particularly useful for analyzing discontinuities, where during a short time period an extensive range of frequencies is present. It can be noted that the instance of glottal closure is represented by a discontinuity in the speech waveform.

The characterization of wavelets centers around the so-called mother wavelet ψ, from which a class of wavelets can be derived as follows [23]:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{2}$$

where a is the frequency or dilation variable and b is the position, or time-domain translation, parameter. Thus, wavelets exist for every combination of a and b. Koornwinder's book [23] is suggested for further augmenting these concepts.

In this paper, the polynomial splines suggested by Mallat and Zhong [20] are adopted for the wavelets. These polynomial spline wavelets can be implemented effectively using a pyramidal algorithm similar to subband filtering, as shown in Fig. 1, where the high-pass-filtered signals are $D_j(w)$, while the low-pass-filtered signals are $A_j(w)$.

The discrete dyadic wavelet transform D_YWT, $\psi_{a,b}(t)$, as defined by Kadambe and Boudreaux-Bartels [18], is exemplified in Fig. 2 for a 20-ms segment of speech. As the wavelet scales increase toward the bottom of the figure, the periodicity of the speech signal becomes more evident from both the time- and frequency-domain plots shown in Fig. 2 for the $D_j(w)$ signals. We note here that although the time-domain waveforms of Fig. 2 are plotted on the finest scale, corresponding to a sampling frequency of 8 kHz, they are waveforms that can be subsampled by a factor of 2^j and thus have lower effective time-domain resolution, as evidenced by their frequency-domain plots. Hence, while the higher wavelet scales give a clear indication of the pitch frequency, the lower scales give the most accurate description of the time-domain position of discontinuities, which are typically associated with the glottal closure instants (GCI's).

The reduction in computational complexity of the ACF is achieved by reducing the number of pitch periods which require their autocorrelation determined, thus, the wavelet analysis determines a set of candidate pitch periods, which are then passed to the ACF process. The selection of these candidate pitch periods is described next.

Observing Fig. 2, we concluded that some form of preprocessing must be performed in order to determine the instants of glottal closure and, hence, the fundamental or pitch frequency

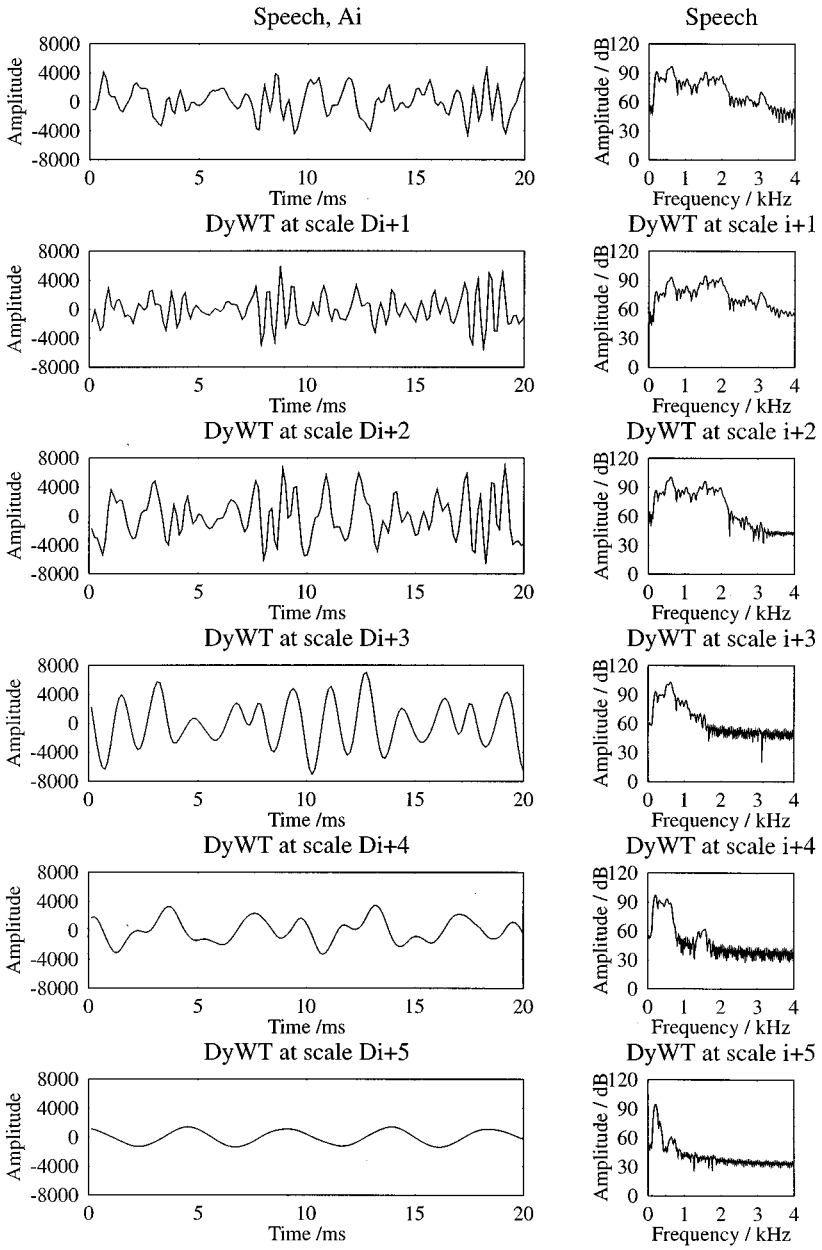

Fig. 2. The D_YWT of 20 ms of speech for a male speaker uttering part of the "i" in "live." For each scale of the D_YWT, the time- and frequency-domain responses are portrayed, enabling the process of the D_YWT to be clearly interpreted. The D_YWT scales are 2000–4000, 1000–2000, 500–1000, 250–500, and 125–250 Hz, respectively, while the sampling rate is 8 kHz.

of the speech waveform. In Fig. 2, the maxima and minima during each scale of the D_YWT provides the most pertinent information about the speech waveform's pitch period. Hence, the left-hand trace of Fig. 3 illustrates the initial preprocessing, whereby positive impulses are placed at the time-domain waveform maxima and negative impulses at the minima. Each of these impulses is assumed to represent possible instants of glottal closure.

The highest permitted fundamental frequency in the context of speech coding is typically 400 Hz, corresponding to a pitch period of 2.5 ms, which imposes limited practical constraints, since most high-pitch female speakers or children have a pitch lower than 400 Hz. Hence, the impulses of Fig. 3(a) placed at the maxima must be at least 2.5 ms apart, and, similarly, the impulses placed at the minima are also at least 2.5 ms apart. Additionally, only impulses which occur in every wavelet scale are considered as potential glottal pulse locations. Finally, the GCI's are normalized and combined, as follows. Each pulse is divided by the largest pulse in that scale, and the pulses across the scales are subsequently added together in order to produce the combined pulse. The impulse magnitudes indicate our confidence in the assumed position of the GCI. This process is characterized in Fig. 3(c) and (d). Assuming that the largest positive and negative pulses are true glottal pulse locations, a range of possible pitch periods can be calculated. Namely, the candidate pitch periods are classified on the basis of the time durations between the largest positive pulse and all other positive pulses, or the largest negative pulse and all other negative pulses.

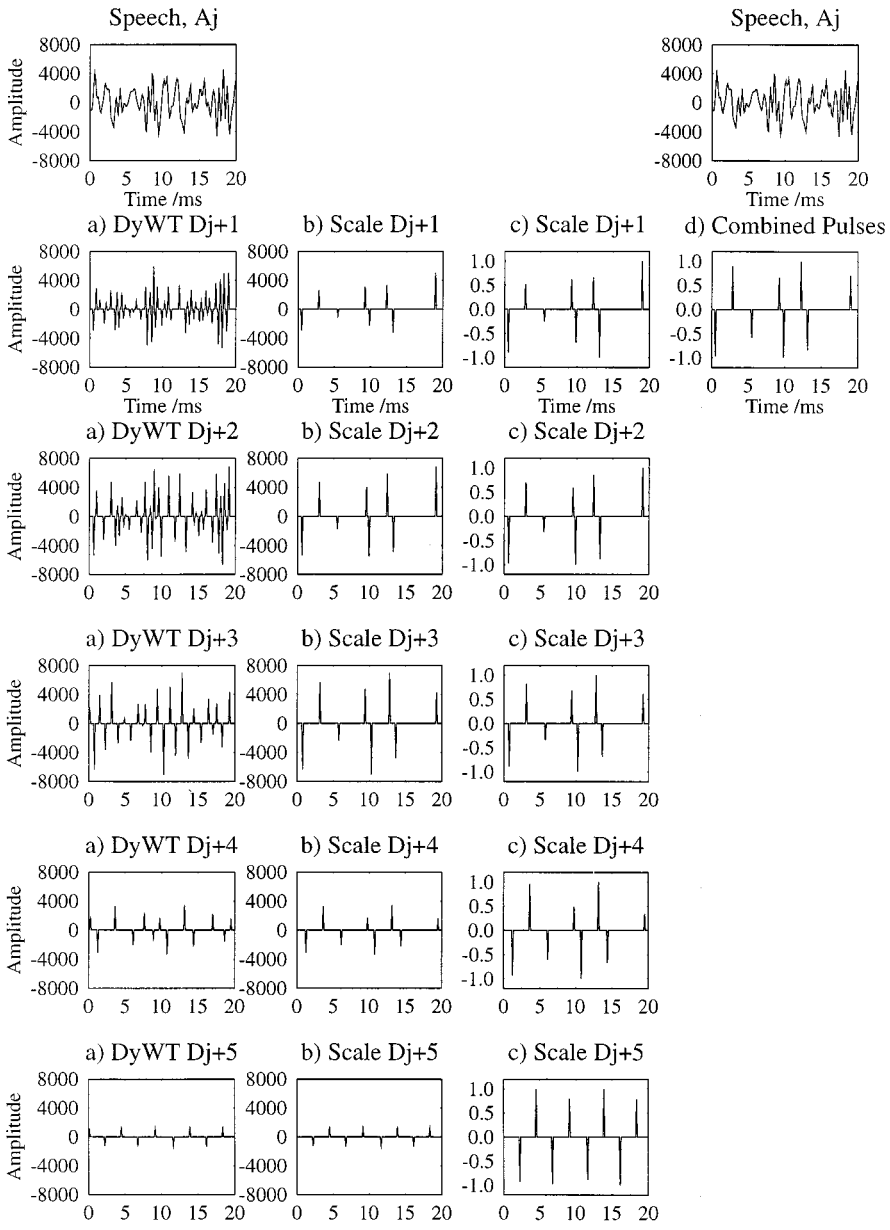

Fig. 3. The D_YWT of 20 ms of speech for a male speaker uttering a fraction of the "i" in "live." In column (a), the corresponding impulses have been placed at the locations of the maxima and minima of the detail signals $D_{j+1} \cdots D_{j+5}$ of Fig. 2. In column (b), only those maxima and minima which persist in every scale are kept. In column (c), all the maxima and minima are normalized with division by the largest pulse in the scale. Finally, in column (d) the scales are added together producing combined maxima and minima representing all scales.

The candidate pitch periods from the D_YWT can now be used to reduce the computational complexity of the ACF. However, first a brief description of how wavelet analysis can be used for V/U decisions is given.

A. Voiced–Unvoiced Decisions

The ability of the D_YWT to categorize speech as voiced or unvoiced has been shown previously [18], [24], and hence similar to these two methods, we briefly describe a V/U decision method based on the energy of the speech signal. The process of the D_YWT across the scales gradually removes the higher frequency components present in the speech waveform, as it was shown in Fig. 2. For unvoiced speech, most energy is present in the higher frequencies, while voiced speech has more of a low-pass nature. A suitable parameter for evaluating V/U decisions was found to be the ratio of the RMS energy in the frequency range 2–4 kHz, to that in the frequency band 0–2 kHz.

B. Pitch Estimation

Fig. 4 displays the potential pitch periods for each speech frame in a speech file. The resultant graph is fairly complex, however, it can be observed that the candidate pitch periods are commonly placed at the true pitch period and its harmonics. Typically, the true pitch period and two or three harmonics are present. There can be at most 15 candidate pitch periods. Namely, in each 20-ms speech frame there can be a maximum of seven pitch intervals that are spaced at least 20 samples apart, for both positive and negative pitch-related pulses.

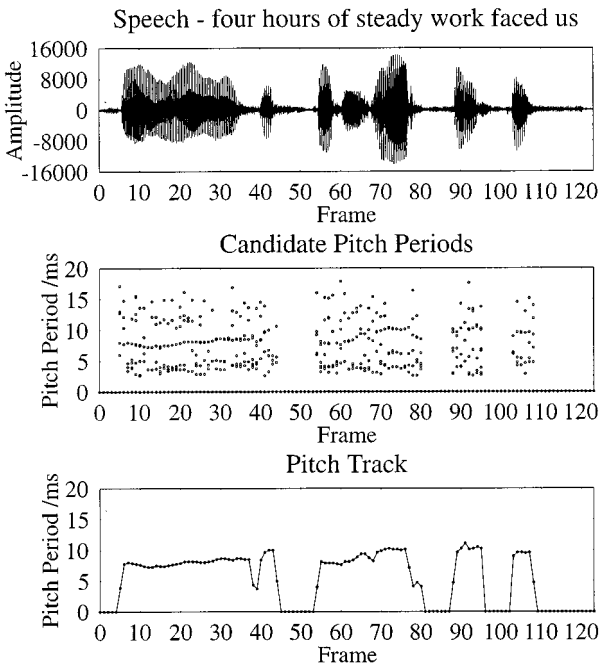

Fig. 4. The original speech signal (top trace), the candidate pitch periods (middle trace) for a British male speaker, and the decided pitch periods (bottom trace) versus LPC frame index for 20-ms frames. The potential pitch periods tend to consist of the pitch frequency track and its harmonics. The detector suffers from pitch halving around frames 40 and 80. According to Table I, the bottom trace was deemed to indicate the "true" pitch with respect to the manually traced pitch 96.1% of the time.

TABLE I
A COMPARISON BETWEEN THE PERFORMANCE OF ACF-BASED PITCH ESTIMATION WITH AND WITHOUT INCORPORATING WAVELET ANALYSIS. W_U REPRESENTS THE PERCENTAGE OF FRAMES THAT ARE LABELED VOICED WHEN THEY SHOULD HAVE BEEN IDENTIFIED AS UNVOICED. W_V INDICATES THE NUMBER OF FRAMES THAT HAVE BEEN LABELED AS UNVOICED WHEN THEY ARE ACTUALLY VOICED. P_G REPRESENTS THE NUMBER OF FRAMES WHERE A GROSS PITCH ERROR HAS OCCURRED. THE TOTAL NUMBER OF INCORRECT FRAMES IS GIVEN AS $W_U + W_V + P_G$

Pitch Detector	W_U%	W_V%	P_G%	Total %
Wavelets-based ACF	1.3	0.3	2.3	3.9
ACF (Inmarsat)	3.1	2.3	1.8	7.2
ACF (G.728)	1.6	5.3	5.8	12.7

Additionally, the previous pitch period may be reintroduced, yielding a total of 15 potential GCI locations, for which the autocorrelation has to be computed.

The performance of the proposed wavelet-based ACF pitch detector was compared to the performance of the conventional "stand-alone" ACF-based pitch detector using pitch tracking according to the G.728 standard [33] without the $D_Y WT$ preprocessing invoked to determine candidate pitch periods. Furthermore, we also compared the proposed wavelet-based pitch detector's performance to that of the Inmarsat [34] in the second line of Table I. The performance of these pitch detectors was evaluated using 20 s of mixed male and female speech, which had been manually pitch tracked. A pitch-estimation error was recorded, when either the V/U detector operated incorrectly or if a gross pitch error occurred. The results are shown in Table I, where it can be seen that the wavelet-based ACF pitch period detector has the lowest overall error rate of $W_U + W_V + P_G$ of 3.9%, which was defined in the caption of Table I.

The performance of the proposed wavelet-based ACF pitch detector and that of the stand-alone G.728 ACF pitch detector can also be compared in terms of computational complexity, which was estimated on the basis of the combined additions and multiplications encountered. For the G.728 ACF pitch detector, the computational complexity is based on the autocorrelation value determined for the legitimate pitch periods of 20–147 samples at a sampling frequency of 8 kHz, or 127 values, producing a computational complexity of 3.35 MFLOPS. The computational complexity of the wavelet-based ACF pitch detector is dependent on two factors, namely, the wavelet analysis and the autocorrelation calculation for the 15 candidate pitch periods. These produced complexities of 2.23 MFLOPS and 0.62 MFLOPS, respectively. Thus, the wavelet-based ACF pitch detector has a lower overall computational complexity of 2.85 MFLOPS, resulting in a 0.5-MFLOP complexity reduction.

Following the above employment of the wavelet transform to reduce the complexity of an autocorrelation-based pitch detector, this pitch detector was included in a waveform interpolation low bit rate speech codec, which is the topic of Section III.

III. BASIC CODING ALGORITHM

Our WI codec of Fig. 5 operates on 20-ms speech frames, for which tenth-order LPC analysis is performed. The LPC coefficients are transformed to line spectrum frequencies (LSF's) and vector quantized to 18 b/frame using an LSF coding scheme similar to that of the G.729 ITU codec [25]. Following LPC analysis, pitch estimation is employed, which is often referred to as pitch detection, followed by a V/U decision, where the pitch-estimation algorithm is based on the novel technique of employing the wavelet transform described above in Section II. For this pitch detector, the pitch period is the distance between two adjacent located GCI's. For an unvoiced frame, the root mean square (RMS) value of the LPC residual is determined, allowing random Gaussian noise to be scaled appropriately and used as unvoiced excitation. The speech waveform is perceptually weighted [26], [27] in order to allocate most of the coding noise in the frequency regions of the speech formants and hence to mask the effects of the coding noise. Additionally, postfiltering [28] is employed on the synthesized speech to improve the perceptual quality.

Since voiced speech segments typically exhibit a higher perceptual importance than unvoiced frames, these segments are more comprehensively defined in our codec, as will be detailed below. For a voiced perceptually weighted speech frame, a prototype segment is selected according to the procedure to be described in Section III-A. This prototype segment represents a full cycle of the pitch period, which is then passed to an analysis-by-synthesis loop, as portrayed in Fig. 5 and detailed throughout the rest of the paper, for the selection of the best voiced excitation. Explicitly, we opted for using the orthogonal zinc basis functions of Fig. 6 in order to model the voiced prototype segments, which, owing to their specific shapes were shown by Sukkar *et al.* [29] to outperform the Fourier series-based representation of the prediction residual in analysis-by-synthesis

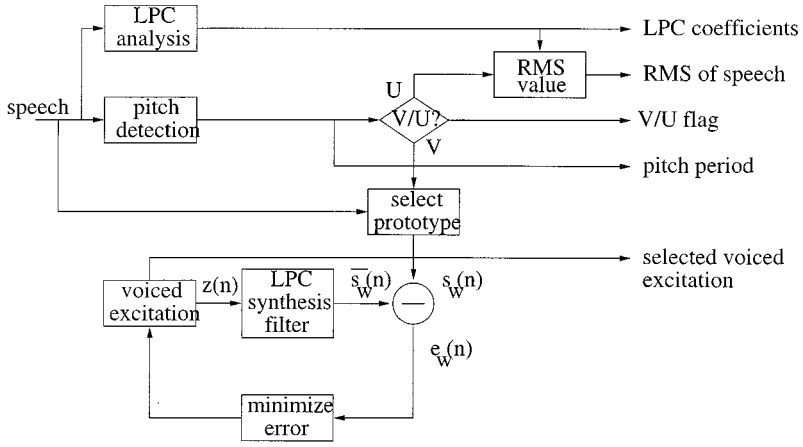

Fig. 5. Schematic of a time-domain prototype WI arrangement.

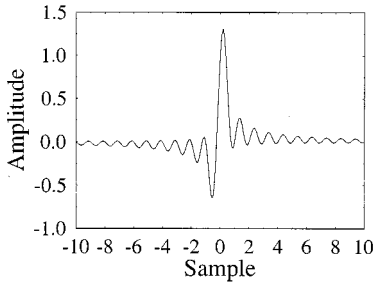

Fig. 6. Stylized shape of a zinc basis function, using the expression $z(n-\lambda) = A \cdot \mathrm{Sin}\, c(n-\lambda) + B \cdot \mathrm{Cos}\, c(n-\lambda)$.

coding of speech. Accordingly, analysis-by-synthesis excitation optimization is invoked, in order to determine the best zinc function excitation (ZFE) for each prototype segment of voiced speech, a technique proposed by Hiotakakos and Xydeas [10]. The proposed speech codec is termed a prototype waveform interpolation zinc function excitation (PWI-ZFE) scheme. These issues will be elaborated on in Section III-B and will also be detailed with reference to the characteristic waveforms of Fig. 8. The ZFE is then quantized and the corresponding parameters detailed in Section III-B are passed to the decoder. At the decoder the excitation is determined by interpolating between the adjacent excitation prototype segments, an issue to be treated in more depth in Section III-F in the context of Fig. 8. Subsequently, the excitation generated by interpolation is passed through the LPC synthesis filter of Fig. 5 in order to reproduce the synthesized speech signal.

Following the above rudimentary overview of the speech codec, the next section offers a more detailed discussion on the different sections of the speech codec shown in Fig. 5. Particular emphasis will be placed on the ZFE optimization process. Let us continue by considering the pitch prototype segment's identification.

A. Pitch Prototype Segment

The determination of the prototype segment in a sequence of voiced speech frames commences by selecting the pitch prototype segment for the first frame in a voiced frame sequence, as it was suggested by Hiotakakos and Xydeas [10]. If P is the pitch of the voiced frame, which was determined by our wavelet-based pitch detector of Section II, then the prototype segment will also be P samples in length. The prototype segment is deemed to begin at a zero crossing immediately to the left of a speech waveform maximum near the center of the speech frame, an approach suggested and detailed by Hiotakakos and Xydeas in [10].

For the subsequent prototype segments, within the voiced sequence of frames the current pitch prototype segment is found by employing Kleijn's cross-correlation-based technique [8]. Explicitly, the prototype segment is located by finding the position of maximum cross correlation between the current speech frame and the previous prototype segment, ensuring as much similarity between prototype segments as possible. Let us now highlight the features of the zinc basis function.

B. Zinc Function Excitation

As mentioned above, the voiced excitations of our codec were derived from the orthogonal zinc basis functions [29], which have previously been advocated by Hiotakakos and Xydeas [10] for a sophisticated higher bit rate interpolation scheme. The zinc function $z(t)$ was defined by Sukkar *et al.* [29] as

$$z(t) = A \cdot \mathrm{Sin}\, c(t) + B \cdot \mathrm{Cos}\, c(t) \tag{3}$$

where

$$\mathrm{Sin}\, c(t) = \frac{\sin(2\pi f_c(t))}{2\pi f_c(t)} \tag{4}$$

and

$$\mathrm{Cos}\, c(t) = \frac{1-\cos(2\pi f_c(t)}{2\pi f_c(t)}. \tag{5}$$

For discrete time processing with a speech bandwidth of $f_c = 4$ kHz and a sampling frequency of $f_s = 2 \cdot f_c = 8$ kHz, we have [10]

$$z(n) = A \cdot \mathrm{Sin}\, c(n) + B \cdot \mathrm{Cos}\, c(n) = \begin{cases} A, & n = 0 \\ \dfrac{2B}{(n)\pi}, & n = \text{odd} \\ 0, & n = \text{even}. \end{cases} \tag{6}$$

A shifted discrete version can be introduced where, $k = \lambda/f_s$ and $t = n/f_s$ and the shifted zinc function is

$$z(n-\lambda) = A \cdot \mathrm{Sin}\, c(n-\lambda) + B \cdot \mathrm{Cos}\, c(n-\lambda) = \begin{cases} A, & n-\lambda{=}0 \\ \dfrac{2B}{(n-\lambda)\pi}, & n-\lambda{=}\text{odd} \\ 0, & n-\lambda = \text{even}. \end{cases} \quad (7)$$

The ZFE model's typical shape is shown by Fig. 6, where the coefficients A and B describe the function's amplitude and λ defines its position.

Once the A and B parameters have been determined, they are Max–Lloyd scalar quantized with 6 b for each A and B parameter.

C. Excitation Optimization

From Fig. 5, the perceptually weighted error signal $e_w(n)$ can be described by [10]

$$e_w(n) = s_w(n) - \overline{s}_w(n) \quad (8)$$

$$= s_w(n) - m(n) - (z(n) * h(n)) \quad (9)$$

where $m(n)$ is the memory of the LPC synthesis filter due to previous excitation segments, while $h(n)$ is the impulse response of the synthesis filter. Thus, the optimization of the excitation signal involves comparing the perceptually weighted error signal $e_w(n)$ for all legitimate values of λ in the range of [1 $\rightarrow$ pitch period] and calculating the corresponding optimum A and B values, which minimize the weighted error for the given λ. The filter memory $m(n)$ was the same as used by Hiotakakos and Xydeas [10].

For the ZFE, there are four possible combinations of positive- or negative-valued A and B parameters, with each of these combinations defined in this contribution using "loose parlance" as a possible phase. If the ZFE pulse polarity defined this way is not maintained throughout a voiced frame sequence, simply because the optimum A or B value has changed sign, the smooth ZFE interpolation process will introduce a sign change for A or B. This results in some small-valued interpolated ZFE's, as the values of A or B pass through zero. For each legitimate zinc pulse position of λ, the sign of A and B is initially checked during the excitation optimization process, and only if the phase restriction of the voiced frame sequence is maintained is the excitation deemed valid. It is feasible that a suitably phased ZFE will not be found. If this occurs, then the previous ZFE is scaled using the RMS value of the LPC residual and repeated for the current voiced frame [10].

D. Complexity Reduction

The complexity of the error minimization process, described in Section III-C, is critical in terms of determining the practicality of the codec. The associated complexity for the optimization is evaluated as follows. The ZFE optimization has a computational complexity dominated by the convolution between the sinc and cosc functions and the impulse response $h(n)$, which is necessary, according to the schematic of Fig. 5, for the optimization loop. This complexity is dependent on the pitch period, or length of the prototype segment, where the complexity dependence on the pitch period is created by the prototype segment length, over which the convolution is performed that may vary from 20 to 147 samples or 50- to 400-Hz fundamental frequency. The dashed line curve of Fig. 7 demonstrates the relationship between the ZFE optimization complexity and pitch period when no restrictions are imposed on this optimization process.

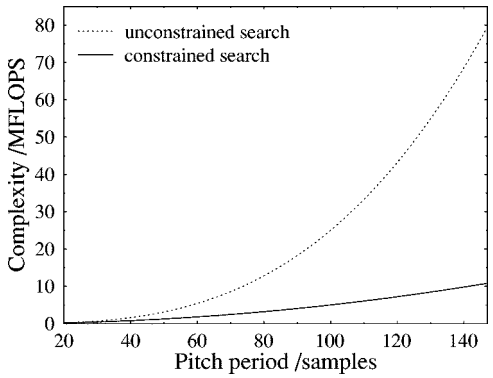

Fig. 7. Computational complexity for the permitted pitch period range of 20–147 sample duration, for both an unrestricted and constrained search.

This curve indicates that if every location λ within the prototype segment were examined, the complexity of ZFE optimization would be prohibitive for real-time implementations. The complexity increase is exponential, as shown by Fig. 7, where it can be seen that any pitch period greater than 90 samples in duration will exceed a complexity of 20 MFLOPS in terms of the ZFE optimization search.

The complexity of the ZFE minimization procedure can be reduced by considering the GCI's introduced in Section II. In Section II, wavelet analysis was harnessed to produce a pitch detector, where the pitch period was determined as the distance between two GCI's. These GCI's indicate the snapping shut, or closure, of the vocal folds, which provides the impetus for the following pitch period. The energy peak caused by the GCI will typically be in close proximity to the position of the ZFE placed by the ZFE optimization process. This allows the complexity reduction of the analysis-by-synthesis process. As noted before, Fig. 7 shows that as the number of possible ZFE positions increases linearly, the computational complexity increases exponentially. Hence, constraining the number of ZFE positions will ensure that the computational complexity remains at a realistic level. This constraining process is described next.

The first frame in a voiced frame sequence invokes no minimization procedure—a single ZFE pulse is simply situated at the glottal pulse location within the prototype segment. For the other voiced frames, in order to maintain a moderate computational complexity, the number of possible ZFE positions is restricted as if the pitch periods were always 20 samples. A suitable constraint is to have the ZFE located within ± 10 samples of the instant of glottal closure situated in the pitch prototype segment. In Fig. 7, the solid line represents the computational complexity of a restricted search procedure in locating the ZFE. The maximum complexity for a 147 sample pitch period is 11 MFLOPS.

We note, however, that constraining the location of the ZFE pulse to within ± 10 positions with respect to the GCI's reduces

TABLE II
SEGSNR RESULTS FOR THE ZFE OPTIMIZATION PROCESS IN VOICED SEGMENTS, WHERE IN CONTRAST TO UNVOICED SEGMENTS THE SEGSNR CAN BE COMPUTED, WITH AND WITHOUT PHASE RESTRICTIONS OR A CONSTRAINED SEARCH. THE SEGSNR RESULTS ARE FOR A MIXTURE OF MALE AND FEMALE SPEAKERS

	unconstrained search	constrained search
no phase restrictions	3.36dB	2.68dB
phase restrictions	2.49dB	1.36dB

the segmental signal-to-noise ratio (SEGSNR) of the weighted prototype voiced speech segments—which was defined as $E[s_w^2(n)]/E[c_w^2(n)]$—which can be seen in Table II. The major drawback of the constrained search is the possibility that the optimization process is degraded through the limited range of ZFE locations searched. Additionally, it is possible to observe a slight degradation to the mean-squared error (MSE) optimization, caused by the phase restrictions imposed on the ZFE's, necessary to permit smooth interpolation. Table II displays the SEGSNR values of the concatenated purely voiced prototype speech segments for which the SEGSNR values can be computed. By contrast, the unvoiced segments are ignored, since these speech segments are represented by noise, thus, a SEGSNR value would be meaningless.

Observing Table II for a totally unconstrained search, the SEGSNR achieved by the ZFE optimization loop is 3.36 dB. The process of either implementing the above-mentioned ZFE phase restriction or constraining the permitted ZFE locations to the vicinity of the GCI's reduces the voiced segments' SEGSNR after ZFE optimization by 0.87 and 0.68 dB, respectively. Restricting both the phase and ZFE locations reduces the SEGSNR by 2 dB. However, in perceptual terms the ZFE interpolation procedure, described in Section III-F, actually improves the subjective quality of the decoded speech due to the smooth speech waveform evolution facilitated, despite the SEGSNR degradation of about 0.87 dB caused by imposing phase restrictions. To assess the impact of constraining the location of the ZFE's, listening tests were conducted, where eight listeners were asked to express a preference between the output speech from constraining the ZFE locations and not constraining the ZFE locations. Three sentences were played to each listener. It was found that 45.8% of listeners preferred the output speech where the ZFE locations had been constrained, while 54.2% of listeners preferred the output speech, where the ZFE locations had not been constrained. The preference values allowed us to justify constraining the ZFE locations, yielding the corresponding computational complexity reductions seen in Fig. 7.

We now proceed by devoting some attention to improving the representation quality of V/U transitions.

E. Voiced Unvoiced Transition

In low bit rate speech codecs, typically the worst represented portion of speech is the rapidly evolving onset of voiced speech. Previous speech codecs have been found to produce better quality speech by locating the emergence of voicing as precisely as possible [10], [30]. Once again, the GCI's inferred from our wavelet transform-based pitch detector of Section II are used to determine the onset of voicing. Specifically, if frame N is voiced and frame $N-1$ is unvoiced, then the end of frame $N-1$ is examined for the evidence of an emerging voiced segment. If GCI's exist at or near the locations, which would maintain the periodicity of voiced speech in frame $N-1$, then the voiced speech region is extended to cover the end of the predominantly unvoiced frame $N-1$, otherwise, the region of speech belonging to frame $N-1$ is confirmed as purely unvoiced. A similar procedure is implemented at the end of a string of voiced frames. We marked the location of the V/U transition by the parameter b_s, which encodes the number of voiced pitch-duration speech cycles within unvoiced frames. Following this description of the speech encoder, the interpolation process harnessed in the decoder is examined.

F. Interpolation

The adopted ZFE parametric representation of the voiced excitation permits simple linear interpolation at the decoder in order to reinsert the zinc pulses at the locations, for which no pulses were transmitted. These issues are detailed below with reference to Fig. 8. Fig. 8 follows the spirit of the work by Hiotakakos and Xydeas [10] and shows an example of the ZFE excitation-based reconstruction of a 60-ms speech segment for a female speaker. Specifically, the top trace shows a 60-ms segment of the original speech signal, the second trace displays the prototype segments identified, while the third one shows the corresponding ZFE. In the fourth trace, the ZFE amplitude parameters A and B are linearly interpolated between the corresponding A and B values of the prototype segments. The bottom trace shows the reproduced speech waveform, which exhibits a close waveform similarity with the original speech segment.

Interpolating the position of the ZFE's linearly—like we did for the amplitudes—will not produce a smoothly evolving excitation and reconstructed speech signal, although speech typically has a smoothly evolving nature. Thus, the most perceptually pleasing output is produced by a smoothly evolving excitation. A smoother output waveform is produced when the pulse position λ is kept constant within each prototype segment throughout a voiced frame sequence. This introduces time misalignment between the original and synthesized waveforms, but produces a smooth excitation signal. The transmission of the λ parameter when it is kept constant contains inherent redundancy as it is transmitted every frame. As an improvement of the scheme proposed in [10], it is suggested that the true position of the ZFE pulse, λ, is arbitrary and hence need not be transmitted. Following this hypothesis, our experience shows that we can set $\lambda = 0$ at the decoder, which has no degrading effect on the speech quality.

It is safe to place λ at the edge of each prototype segment since every ZFE is permitted to extend over three interpolation regions, namely, its allotted region together with the previous and the next region. This allows ZFE's at the interpolation region boundaries to be fully represented in the excitation waveform by ensuring that every ZFE will have a tapered low energy value when it is curtailed.

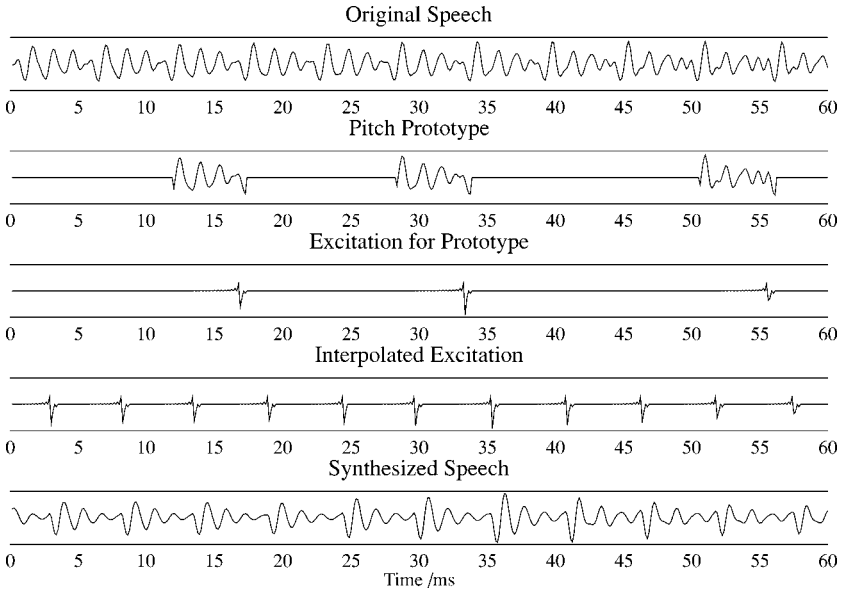

Fig. 8. Example of 60-ms segments of the original speech (top to bottom), the pitch prototype, and its zinc model as well as the interpolated excitation and the synthesized speech for a voiced utterance by a female speaker uttering /*/ in "dog."

In addition to the ZFE amplitude and position parameter, the decoder interpolation procedure requires the zero-crossing parameter of the prototype segments, which was introduced in [10]. However, the zero-crossing values of the prototype segments are typically a frame length apart, since a prototype segment is selected once every frame. Hence, instead of explicitly transmitting the zero-crossing parameter, it can be assumed that the start of the consecutive prototype segments is a frame length apart. An arbitrary starting point for the prototype segments could be $FL/2$, where FL is the speech frame length.

G. 1.9-kbps ZFE-WI Codec Performance

The ZFE-WI codec performance was considered for both male and female speakers. Here, a female speaker is given as an example. The speech segment displayed in Fig. 9(a) was recorded for a female speaker. As seen by comparing Fig. 9(a) and (c) in the time domain, the basic shape of the original waveform is more or less preserved—apart from an arbitrary phase shift, which is in this case π. Additionally, by observing Fig. 9(a) and (c), we note from the frequency-domain plots that the formant location is preserved, however, it is noticeable that the inclusion of unvoiced speech above 1800 Hz is not modeled well by the distinct V/U nature of the PWI-ZFE scheme. Observing the ZFE waveform of Fig. 9(b), a flat excitation frequency-domain envelope is produced, while its spectral fine structure reflects the pitch-dependent needle-like behavior. Informal listening tests showed that the reproduced speech contained slight "buzziness," and hence its quality was deemed inferior in comparison to the original speech.

The bit allocation for the ZFE coder is summarized in Table III, where 18 b are reserved for LSF vector quantization [25], while a 1-b flag is used for the V/U classifier. For unvoiced speech the RMS parameter is scalar quantized with 5 b. The b_s offset requires a maximum of 3 b to encode the V/U transition point in terms of the number of voiced speech cycles within unvoiced frames, since assuming a minimum pitch duration of 20 samples, a maximum of eight pitch cycles can fit in a 160-sample speech frame. For voiced speech the pitch can vary from 20 to 147 samples, thus requiring 7 b for transmission. The ZFE amplitude parameters A and B are scalar quantized with 6 b. Following the above investigations of the proposed PWI-ZFE speech codec, we now invoke mixed multiband excitation (MMBE) in order to improve the quality of the speech codec, while increasing the bit rate from 1.9 to 2.4 kbps.

IV. Multiband Excited Codec

Speech typically contains a mixture of voiced and unvoiced excitation across its frequency bandwidth. Thus, the division of speech into V/U frames, which has been performed so far, does not follow the true nature of the speech signal. The well-known speech coding technique of multiband excitation (MBE) [9], which is briefly explained next, is capable of allowing a mixture of voiced and unvoiced excitation in each speech frame.

A. The MMBE Coding Algorithm

The encoder and decoder schematics of a MMBE architecture are shown in Fig. 10, where following short-term predictive (STP) LPC analysis of the 20-ms speech frame, pitch estimation is invoked in order to locate any evidence of voicing. A frame deemed unvoiced has the root-mean-squared value of its LPC residual quantized and is sent to the decoder.

Speech frames labeled as voiced are split into M frequency bands, with M constrained to be a time-invariant constant value. In our scheme, $M = 3$ bands were used. Each of the $M = 3$ frequency bands is examined for evidence of voicing [31] and has a voicing strength assigned that was scalar quantized to eight different strengths, allowing us to assign a total of $3 \times 3 = 9$ b per 20 ms to the three bands. Hence, a total rate of 0.45 kbps was required for voicing strength quantization. The 4-kHz frequency spectrum was divided into three frequency bands depending on the pitch of the speech frame [32]. The voiced excitation must

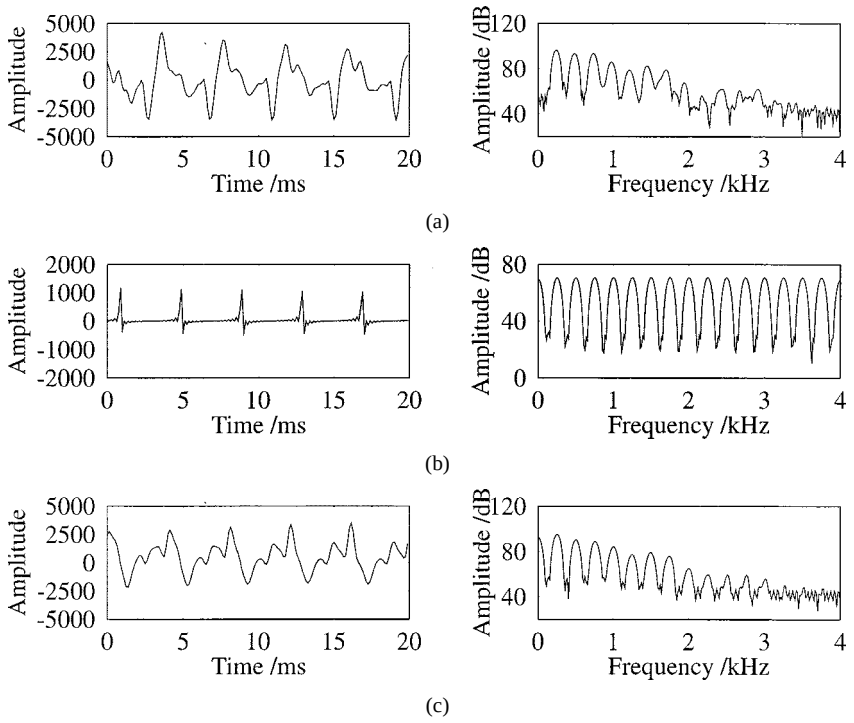

Fig. 9. Time- and frequency-domain comparison of the (a) original speech, (b) ZFE waveform, and (c) output speech after the pulse dispersion filter. The 20-ms speech frame is the liquid /r/ in the utterance "*r*ice" for a female speaker.

TABLE III
BIT ALLOCATION OF THE 1.9-kbps PWI-ZFE SPEECH CODEC

parameter	unvoiced	voiced
LSFs	18	18
v/u flag	1	1
RMS value	5	-
b_s	3	-
pitch	-	7
A	-	6
B	-	6
total/20ms	27	38
bit rate	1.35kbps	1.90kbps

also be determined and its parameters sent to the decoder. In our proposed scheme, the previously described PWI-ZFE speech codec model was employed, again invoking the classic perceptual weighting principles of the previous described PWI-ZFE speech codec.

As seen in Fig. 10, at the decoder both unvoiced and voiced speech frames have an associated pair of filter banks created [31]. A consequence of the time-variant pitch period is the need for the filterbank to be reconstructed every frame, in both the encoder and decoder, as shown in Fig. 10, thus increasing the computational costs. However, for unvoiced frames, the filterbank excitation is declared fully unvoiced, and, hence, no voiced excitation is created.

Following Fig. 10, both the voiced and unvoiced decoder filter banks are created using the knowledge of the pitch period and the number of frequency bands M. Specifically, the filter bandwidths have to be an integer multiple of the pitch [9]. For the voiced filter banks, the filter coefficients are scaled by the quantized voicing strengths determined at the encoder. A value of one represents full voicing, while a value of zero signifies a frequency band of noise. Intermediate values represent a mixed excitation source. For the unvoiced filter bank, the voicing strengths are adjusted, ensuring that the voicing strengths of each voiced and unvoiced frequency band combine to unity. This constraint ensures that the combined signal from the filter banks is spectrally flat over the entire frequency range. The mixed excitation speech is then synthesized, as shown in Fig. 10, where the LPC filter determines the spectral envelope of the speech signal.

V. 2.35-kbps ZFE-MMBE-WI-CODED PERFORMANCE

The combined 2.35-kbps three-band PWI-ZFE-MMBE scheme was studied in terms of speech quality for both male and female speakers. The speech frame examined in Fig. 11 is the same utterance as that characterized in Fig. 9. In the time domain, the basic shape of the original waveform is preserved with a phase shift of π. Observing Fig. 11(b) above 2 kHz, a mixture of voiced and unvoiced excitation is harnessed. From Fig. 11(c), it can be seen that the presence of noise above 2 kHz produces a better representation of the frequency spectrum than Fig. 9(c).

Listening tests were conducted to assess the performance of the developed speech coders. Pairwise-comparison tests were performed where eight listeners were played three different sentences from a mixture off male and female speakers. For each different sentence the listeners were played two versions: version A and version B. Having being played each version twice, they were asked to express a preference for version A or version B. The 1.9-kbps PWI-ZFE speech coder was compared with the 2.35-kbps speech coders where MMBE has been added for three frequency bands. From Table IV, it can be seen that 95.8% of

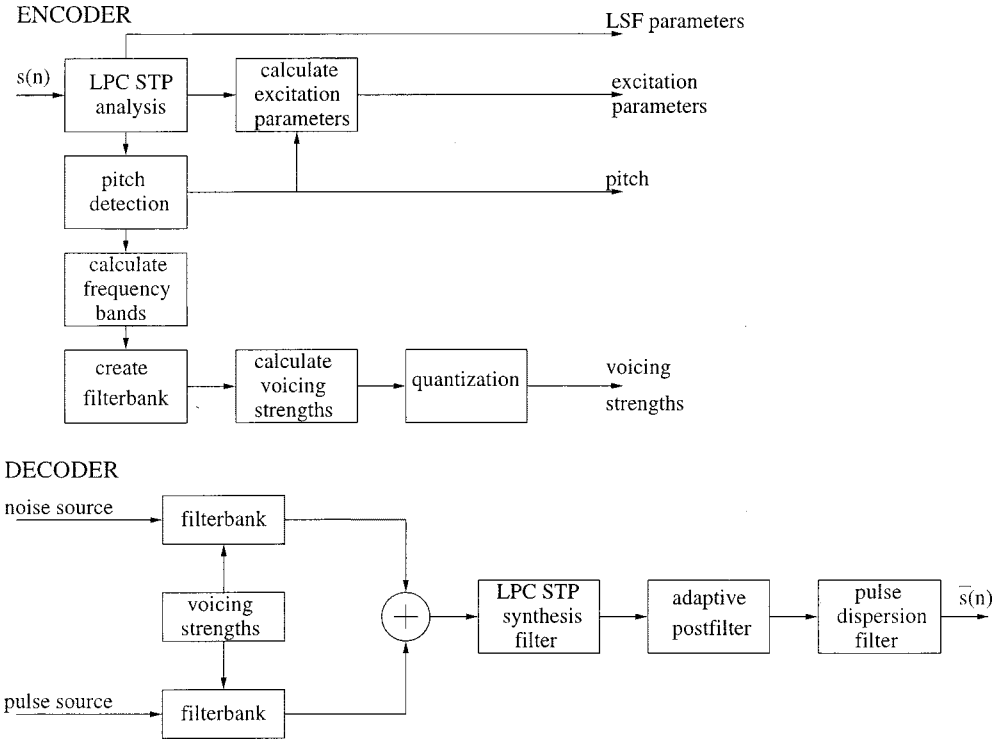

Fig. 10. Schematic of the MMBE encoder and decoder.

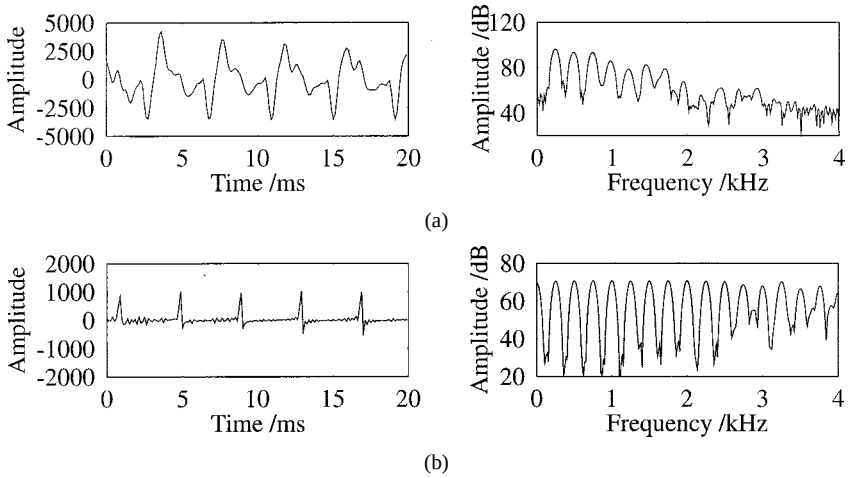

Fig. 11. Time- and frequency-domain comparison of the (a) original speech, (b) three-band MMBE ZFE waveform, and (c) output speech after the so-called pulse dispersion filter. The 20-ms speech frame is the liquid /r/ in the utterance "*r*ice." For comparison with the full-band process, refer to Fig. 9.

listeners preferred the 2.35-kbps speech coder, noting that this speech quality improvement came at the cost of a higher bit rate. The 2.35-kbps three-band MMBE speech coder with PWI-ZFE was also compared with a 2.35-kbps five-band MMBE speech coder where a single pulse was used to represent the voiced excitation. From Table IV, it can be seen that 79.2% of listeners preferred the PWI-ZFE for representing the voiced excitation.

TABLE IV
LISTENING TESTS

Speech Coder A	Speech Coder B	Prefer A%	Prefer B%
1.9kbps PWI-ZFE	2.35kbps 3-band MMBE with PWI-ZFE	4.2	95.8
2.35kbps 3-band MMBE with PWI-ZFE	2.35kbps 5-band MMBE with a single pulse	79.2	20.8

VI. SUMMARY AND CONCLUSIONS

The proposed wavelet-based techniques substantially reduced the pitch-search complexity of our codec. The refined PWI-ZFE codec reduced the bit rate of the scheme proposed in [10], but due to the binary V/U classification it exhibited some "buzziness," which was mitigated by introducing an eight-level voicing strength in each of the three subbands of the MBE-ZFE codec, resulting in a 2.35-kbps arrangement. Our future work is targeted at creating sinusoidally excited benchmarkers.

REFERENCES

[1] D. O'Shaughnessy, *Speech Communication: Human and Machine*. Reading, MA: Addison-Wesley, 1987.

[2] S. Furui, *Digital Speech Processing, Synthesis and Recognition.* New York: Marcel Dekker, 1989.

[3] A. M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communications Systems.* New York: Wiley, 1994.

[4] W. B. Kleijn and J. Haagen, "A speech coder based on decomposition of characteristic waveforms," in *Proc. ICASSP'95*, pp. 508–511.

[5] A. Gersho, "Advances in speech and audio compression," *Proc. IEEE*, pp. 900–918, June 1994.

[6] I. A. Gerson, M. A. Jasiuk, J.-M. Muller, J. M. Nowack, and E. H. Winter, "Speech and channel coding for the half-rate GSM channel," in *Proc. ITG-Fachbericht 130*, Berlin, Germany, Nov. 1994, pp. 225–233.

[7] T. Ohya, H. Suda, and T. Miki, "3.45 kbits/s PSI-CELP of the half-rate PDC speech coding standard," in *Proc. IEEE Conf. Vehicular Technology*, June 1994, pp. 1680–1684.

[8] W. B. Kleijn, "Encoding speech using prototype waveforms," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 386–399, Oct. 1993.

[9] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 8, pp. 1223–1235, 1988.

[10] D. J. Hiotakakos and C. S. Xydeas, "Low bit rate coding using an interpolated zinc excitation model," in *Proc. ICCS'94*, pp. 865–869.

[11] K. A. Teague, B. Leach, and W. Andrews, "Development of a high-quality MBE based vocoder for implementation at 2400bps," in *Proc. IEEE Wichita Conf. Communications, Networking and Signal Processing*, April 1994, pp. 129–133.

[12] H. Hassanein, A. Brind'Amour, S. Dèry, and K. Bryden, "Frequency selective harmonic coding at 2400bps," in *Proc. 37th Midwest Symp. Circuits and Systems*, vol. 2, 1995, pp. 1436–1439.

[13] R. J. McAulay and T. F. Quatieri, "The application of subband coding to improve quality and robustness of the sinusoidal transform coder," in *Proc. ICASSP'93*, vol. 2, pp. 439–442.

[14] A. V. McCree and T. P. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp. 242–250, 1995.

[15] P. A. Laurent and P. de La Noue, "A robust 2400bps subband LPC vocoder," in *Proc. ICASSP'95*, pp. 500–503.

[16] A. V. McCree, K. Truong, E. B. George, T. P. Barnwell, and V. Viswanathan, "A 2.4kbit/s coder candidate for the new U.S. Federal standard," in *Proc. ICASSP'96*, pp. 200–203.

[17] Y. Hiwasaki and K. Mano, "A new 2-kbit/s speech coder based on normalized pitch waveform," in *Proc. ICASSP'97*, pp. 1583–1586.

[18] S. Kadambe and G. F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE Trans. Inform. Theory*, vol. 38, pp. 917–924, March 1992.

[19] N. Gonzalez and D. Docampo, "Application of singularity detection with wavelets for pitch estimation of speech signals," *Signal Processing VII: Theories and Applications*, pp. 1657–1660, 1994.

[20] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, pp. 710–732, July 1992.

[21] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inform. Theory*, vol. 36, pp. 961–1005, September 1990.

[22] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, pp. 674–693, July 1989.

[23] T. H. Koornwinder, "Wavelets: An elementary treatment of theory and applications," *World Scientific*, 1993.

[24] J. Stegmann, G. Scröder, and K. A. Fischer, "Robust classification of speech based on the dyadic wavelet transform with application to CELP coding," in *Proc. ICASSP 96*, pp. 546–549.

[25] "Coding of speech at 8 kbit/s using conjugate-structure algebraic CELP G.729," CCITT Rep., Dec. 1995.

[26] B. S. Atal and M. R. Schroeder, "Predictive coding of speech and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, pp. 247–254, June 1979.

[27] ——, "Optimizing predictive coders for minimum audible noise," in *Proc. ICASSP'79*, pp. 453–455.

[28] J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 59–70, 1995.

[29] R. A. Sukkar, J. L. LoCicero, and J. W. Picone, "Decomposition of the LPC excitation using the zinc basis functions," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 9, pp. 1329–1341, 1989.

[30] K. Yaghmaie and A. M. Kondoz, "Multiband prototype waveform analysis synthesis for very low bit rate speech coding," in *Proc. ICASSP'97*, pp. 1571–1574.

[31] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Keijn and K. K. Paliwal, Eds. New York: Elsevier Science, 1995, ch. 4.

[32] S. Yeldener, A. M. Kondoz, and B. G. Evans, "Multiband linear predictive speech coding at very low bit rates," *Proc. Inst. Elect. Eng.*, vol. 141, pp. 284–296, Oct. 1994.

[33] "Coding of speech at 16 kbit/s using low-delay code excited linear prediction G.728," CCITT Rep., Sept. 1992.

[34] "Inmarsat-M voice codec," DVSI Rep., Aug. 1991.

F. C. A. Brooks (S'96–M'99) received the M.Eng. and Ph.D. degrees in electronic engineering in 1995 and 1999, respectively, from the University of Southampton, Southampton, U.K.

From 1995 to 1998, she performed research on low bit rate speech coders for wireless communication. She is now with the Global Wireless Systems Research Department, Bell Laboratories, Swindon, U.K. Her current research involves the transmission of voice over a packet air interface.

Lajos Hanzo (SM'92) held various research and academic posts in Hungary, Germany, and the United Kingdom during his 23-year career in telecommunications. Since 1986, he has been a Member of the Academic Staff, Department of Electronics and Computer Science, University of Southampton, Southampton, U.K. He has also been a Consultant to Multiple Access Communications Ltd., U.K. As a member of two multinational consortia and funded by the European Community as well as the Engineering and Physical Sciences Research Council (EPSRC), U.K., he is currently conducting research towards the next generation of wireless multimedia systems. He has published widely in *Wireless Multimedia Communications*, including three monographs and in excess of 200 research papers.

Dr. Hanzo currently holds a Chair in Telecommunications. He has organized and chaired conference sessions, presented overview lectures, and was awarded a number of distinctions. He is a member of the IEE. For further information, visit: http://www-mobile.ecs.soton.ac.uk.