A 1.9 KBPS ZINC FUNCTION EXCITED, WAVEFORM INTERPOLATED SPEECH CODEC

F.C.A.Brooks, L. Hanzo

Dept. of Electr. and Comp. Sc.,Univ. of Southampton, SO17 1BJ, UK.
Tel: +44-703-593 125, Fax: +44-703-594 508
Email: lh@ecs.soton.ac.uk http://www-mobile.ecs.soton.ac.uk

ABSTRACT

The Zinc-function excited codecs of References [3, 7] were further developed by including a novel wavelet-based pitch-detector, by reducing its complexity and by eliminating the need for transmitting the Zinc-function positions. The bit allocation of Table 1 summarizes the most salient codec features.

1. INTRODUCTION

The well-established low bit rate speech compression technique of waveform interpolation (WI) was pioneered by Kleijn [1]. WI techniques can be sub-divided into time- and frequency-domain interpolation. In waveform interpolation a characteristic waveform, which is also referred to as a prototype waveform, is periodically located in the original speech signal. Between these selected prototype segments 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 the above-mentioned two sub-classes. The fundamental aim of interpolation-based coders is, hence, to represent a small portion of the waveform, namely the prototype segment accurately, and then perform interpolation between these segments to reproduce the synthesized speech signal. Since only the prototype segments have to be encoded, the required bit rate is low, while maintaining good perceptual speech quality.

Most WI systems rely on frequency domain coding [1, 2], although there are schemes, such as the proposed one, which employ time-domain coding [3, 4]. 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.

The outline of the paper is as follows. In Section 2 we provide an overview of the coding algorithm, followed by the portrayal of the Zinc function excitation (ZFE) [7] for representing voiced speech segments in Section 3. Particular attention is paid to the optimization of the ZFE excitation and to the associated complexity. This is followed by a discussion on further WI operations, such as the selection of the most suitable prototype segment, the significance of the voiced-unvoiced transition point and the implementation of interpolation at the decoder. Finally, the performance of the described codec is considered in Section 4, before concluding in Section 5.

2. CODING ALGORITHM

Our WI codec of Figure 1 operates on 20ms speech frames, for which LPC analysis is performed. The LPC coefficients are transformed to line spectrum frequencies (LSFs) and vector quantized to 18bits/frame using an LSF coding scheme similar to that of the G.729 ITU codec[5]. Following LPC analysis, pitch detection and a voiced-unvoiced (V/U) decision are performed, where the pitch-detection algorithm is based on a novel technique employing the wavelet transform [6]. For this pitch detector the pitch period is the distance between two located glottal closure instants (GCIs). 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.

Due to the human ear's increased sensitivity to voiced speech, these segments are more comprehensively defined, as it will be detailed below. For a voiced speech frame a prototype segment is selected, representing a full cycle of the pitch period. Subsequently the prototype segment is passed to an analysis-by-synthesis loop, as portrayed in Figure 1, in order to select the best voiced excitation. Explicitly, we opted for using orthogonal Zinc basis functions in order to model the prototype segments, which, owing to their specific shapes were shown by Sukkar, Cicero and Picone [7] to outperform the Fourier-transform in analysis-by-synthesis coding of speech. These Zinc basis functions are passed to the analysis-by-synthesis loop, in order to determine the best Zinc function excitation (ZFE) for each prototype segment of voiced speech, a technique proposed by Hiotakakos and Xydeas [3]. They are then quantized and the corresponding parameters are passed to the decoder. At the decoder the excitation for each prototype segment is determined by interpolating between the adjacent segments and subsequently the excitation recovered by interpolation is passed through the LPC synthesis filter in order to reproduce the synthesized speech signal.

GLOBECOM'98, Sydney, Australia

The financial support of the following organizations is gratefully acknowledged: Motorola ECID, Swindon, UK; European Community, Brussels, Belgium; Engineering and Physical Sciences Research Council, Swindon, UK; Mobile Virtual Centre of Excellence, UK.

0-7803-4984-9/98/$10.00 ©1998 IEEE.

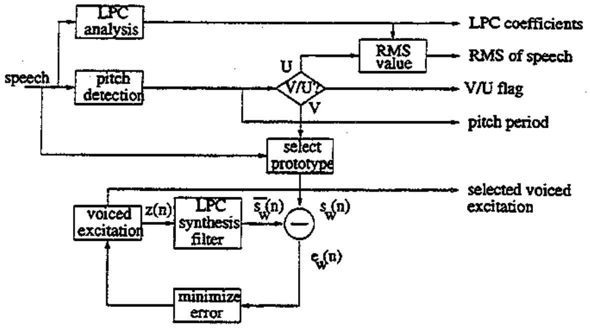

Figure 1: Schematic of a time domain prototype WI system.

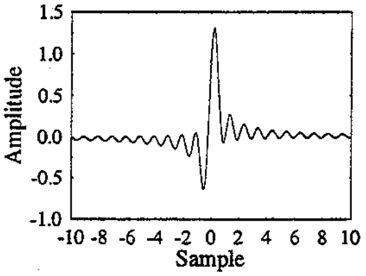

Figure 2: Typical shape of a Zinc basis function, using the expression $z(n) = A \cdot sinc(n - \lambda) + B \cdot cosc(n - \lambda)$.

Following the above rudimentary overview of the speech coder, in the next Section a more detailed discussion is offered on the optimum choice of the ZFE.

3. ZINC FUNCTION EXCITATION

As mentioned above, the voiced excitations of our codec were derived from the orthogonal Zinc basis functions [7], which have previously been advocated by Hiotakakos and Xydeas [3] for a sophisticated higher bit rate interpolation scheme. The Zinc function $z(t)$ was defined by Sukkar et al [7] as:

$$z(t) = A \cdot sinc(t - \lambda) + B \cdot cosc(t - \lambda) \tag{1}$$

where

$$sinc(t) = \frac{\sin(2\pi f_c(t - \lambda))}{2\pi f_c(t - \lambda)} \tag{2}$$

and

$$cosc(t) = \frac{1 - \cos(2\pi f_c(t - \lambda))}{2\pi f_c(t - \lambda)}. \tag{3}$$

For the discrete time case with a speech bandwidth of $f_c = 4kHz$ and a sampling frequency of $f_s = 8kHz$ we have [3]:

$$\begin{aligned} z(n) &= A \cdot sinc(n - \lambda) + B \cdot cosc(n - \lambda) \qquad (4) \\ &= \begin{cases} A & n - \lambda{=}0 \\ \frac{2B}{(n-\lambda)\pi} & n - \lambda{=}\text{odd} \\ 0 & n - \lambda{=}\text{even} \end{cases} \end{aligned}$$

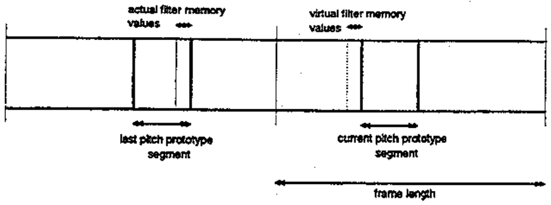
actual filter memory values
virtual filter memory values
last pitch prototype segment
current pitch prototype segment
frame length

Figure 3: Determining the LPC filter memory

The ZFE model's typical shape is shown by Figure 2, where the coefficients A and B describe the function's amplitude and λ defines its position. Sukkar et al [7] compared the Zinc functions to other excitation models, notably the Fourier series. They found the Zinc functions to be superior at modelling the LPC residual, which is due to their pulse-like shape being able to mimic the periodicity of voiced speech that remains after LPC analysis. Furthermore, the presence of energy spread further away from the dominant pulse decreases the synthetic perceptual quality of the speech signal.

3.1. Excitation Optimization

From Figure 1 the error signal $e_w(n)$ can be described by [3]:

$$\begin{aligned} e_w(n) &= s_w(n) - \bar{s}_w(n) & (5) \\ &= s_w(n) - m(n) - (z(n) * h(n)) & (6) \end{aligned}$$

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 error signal $e_w(n)$ for all legitimate values of λ in the range of [1 → pitch period], and calculating corresponding A and B values which minimize the weighted error for the given λ.

The use of prototype segments results in a ZFE determination process that is a discontinuous task, thus the actual filter memory, $m(n)$, is not explicitly available for the ZFE optimization process. Hence, the filter's memory is assumed to be identical to that due to the previous ZFE [3]. Figure 3 shows two consecutive speech frames, where the previous pitch prototype segment has its last p samples highlighted as LPC synthesis filter memory values. For the current pitch prototype segment these p samples have become virtual filter memory. Thus, for the error minimization procedure the speech between the prototype segments has been removed.

There are four possible phases of the ZFE, produced by four combinations of positive or negative valued A and B parameters. If the ZFE phase defined this way is not maintained throughout a voiced sequence the interpolation process will introduce a sign change for A or B, this will result in some small valued interpolated ZFEs, as the values of A or B pass through zero. For each legitimate Zinc pulse position of λ, the sign of A and B are initially checked, and only if the phase restriction of the voiced 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

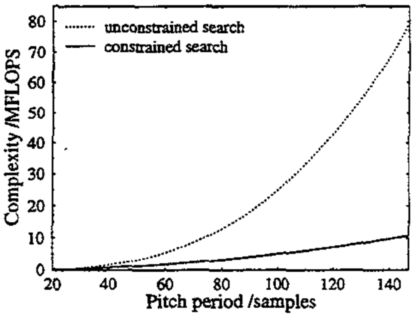

Figure 4: Computational complexity for the permitted pitch period range of 20 to 147 sample duration, for both an unrestricted and constrained search

previous ZFE is scaled using the RMS value of the LPC residual and repeated for the current frame.

3.2. Complexity Reduction

The complexity of the error minimization process described in Section 3.1 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 Figure 1 in order to compute the synthesized speech. This complexity is dependent on the pitch period, or length of the prototype segment. The complexity dependence on the pitch period is created by the prototype segment length, over which the convolution is performed, which may vary from 20 to 147 samples or 50Hz to 400Hz fundamental frequency. The dashed line curve of Figure 4 demonstrates the relationship between the complexity and pitch period, when no restrictions are imposed on this optimization process.

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 Figure 4, 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 pitch-search.

Simplification of this process can be achieved exploiting the knowledge of the Glottal Closure Instants (GCI) located by the wavelet based pitch detector [6] used. The GCI indicates the commencement of the voiced speech cycle and hence it is a likely position for the location of the optimum ZFE. The ZFE optimization complexity can be reduced by constraining the number of ZFE positions λ to the vicinity of these instants, ensuring that the computational complexity remains at a realistic level. A suitable constraint is to have the ZFE located within 10 samples of the instant of glottal closure situated within the pitch prototype segment. The solid line of Figure 4 displays the computational complexity for a restricted search procedure in locating the ZFE. The maximum associated complexity, for a 147 sample pitch period is 10 MFLOPS.

The major drawback of the constrained search is the possibility that the optimization process is severely degraded through the limited range of ZFE locations searched. Additionally it is possible to observe the optimization degradation caused by the phase restrictions imposed on the ZFE to permit smooth interpolation. Table 1 displays the SEGSNR values of the concatenated voiced prototype speech segments. The unvoiced segments are ignored, since these speech spurts are represented by noise, thus a SEGSNR value would be meaningless.

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

Table 1: SEGSNR results for the optimization process with and without phase restrictions, or a constrained search.

Observing Table 1 for a totally unconstrained search, the SEGSNR achieved by the ZFE optimization loop is 3.36dB. The process of either implementing the above-mentioned phase restriction or constraining the permitted ZFE locations to the vicinity of the GCIs reduces the voiced segments' SEGSNR after ZFE optimization by 0.87dB and 0.68dB, respectively. Restricting both the phase and the ZFE locations reduces the SEGSNR by 2dB. However, in perceptual terms the ZFE interpolation procedure implemented actually improves the subjective quality of the decoded speech due to the smooth speech waveform evolution it facilitates, despite the degradation of about 0.87dB caused by imposing phase restrictions. Similarly, the extra degradation of about 1.13dB caused by constraining the location of the ZFEs also improves the perceived decoded speech quality due to smoother waveform interpolation.

3.3. Interpolation Example

Following the spirit of Reference [3] by Hiotakakos and Xydeas, an example of the ZFE excitation based reconstruction of a 60 ms speech segment is demonstrated in Figure 5 for a female speaker. Initially a pitch prototype segment is selected for each of the 20 ms segments and at the encoder the specific ZFE position and A and B parameters minimizing the perceptually weighted error are selected to represent this prototype segment. At the decoder, the ZFE segments between those corresponding to the prototype segments are regenerated by interpolation in order to produce a smoothly evolving excitation waveform. Subsequently this interpolated excitation pattern is passed through the LPC STP synthesis filter to reconstruct the original speech. When constructing the excitation waveform, every ZFE is permitted to extend over three interpolation regions, namely its allotted region together with the previous and future regions. This allows ZFEs near the interpolation region boundaries to be fully represented in the excitation waveform, ensuring that every ZFE will have a low energy value when it is curtailed. During our forthcoming discourse it is beneficial to refer frequently to Figure 5.

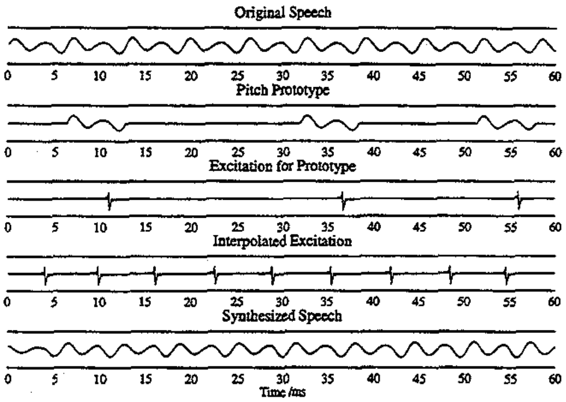

Figure 5: Example of 60 ms segments of the original and synthesized speech, the pitch prototype and its Zinc-model as well as interpolated excitation for a voiced utterance by a female speaker.

3.4. Pitch Prototype Segment

The determination of the first prototype segment in a voiced sequence of speech frames is demonstrated by Figure 6 [3]. If P is the pitch of the voiced frame, which was determined by our wavelet-based pitch-detector [6], then P samples in the centre of the frame are selected as the initial prototype, which is shown at the top of Figure 6. Following the approach proposed by Hiotakakos and Xydeas [3], the actual pitch prototype segment is then deemed to commence at the zero-crossing immediately to the left of the maximum point in the initial prototype selection, as shown in the bottom two graphs of Figure 6. The duration of the pitch prototype segment is P samples.

Locating the start of the first pitch prototype segment near a zero crossing helps to reduce discontinuities in the speech encoding process, resulting in a seamless speech waveform. The other prototypes within the voiced sequence are found by employing Kleijn's cross-correlation based technique [1] where the position of maximum cross correlation between the current speech frame and the previous prototype segment determines the current prototype segment. Then an interpolation process is invoked at the decoder between the consecutive prototype segments, in order to insert the pitch-spaced, linearly amplitude-scaled Zinc basis function excitation components, as demonstrated in the fourth trace of Figure 5.

3.5. 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 [3] [8]. Once again, the GCIs inferred from the wavelet transform based pitch detector [6] 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 GCIs exist at or near the locations, which would maintain the periodicity of voiced speech, then the voiced speech region is extended, otherwise the region of speech is confirmed as unvoiced. A similar procedure is implemented at the end of the voiced region. The location of the voiced-unvoiced transition is represented by the *transition* parameter, which encodes the number of voiced speech cycles within unvoiced frames. The significance of this and the other previously introduced parameters will become more explicit in the context of the bit allocation scheme of Table 2.

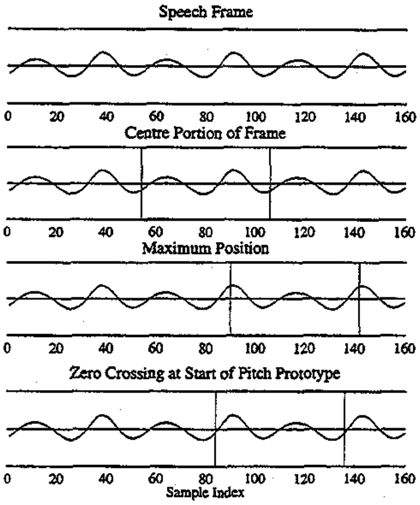

Figure 6: Pitch prototype selection for a female speaker.

3.6. Interpolation

The simple parametric representation of the voiced excitation by the Zinc basis functions guarantees a low bit rate contribution by the excitation encoding. At the decoder seamless interpolation is used between the prototype excitation segments in order to reinsert the Zinc-pulses, which were not transmitted. These issues are detailed below with reference to Figure 5. Specifically, the Zinc function amplitude parameters A and B are linearly interpolated between the corresponding values of the prototype segments, as it is demonstrated in the Figure. The process of cross-correlation based prototype selection technique of Section 3.4 and the choice of restricted locations in the vicinity of the GCIs ensures that the consecutive λ values are similar in adjacent prototype frames. For the sake of smooth speech waveform evolution the Zinc-pulse locations λ are kept constant throughout a voiced speech spurt with respect to the prototype segments.

Recall that the optimum Zinc-pulse position λ was determined relative to the GCIs at the encoder on the basis of finding the location, where the optimum A and B parameters minimized the perceptually weighted error over the duration of the prototype segment of P samples. However, once the optimum A and B parameters are determined, the decoder can reconstruct the synthetic speech without

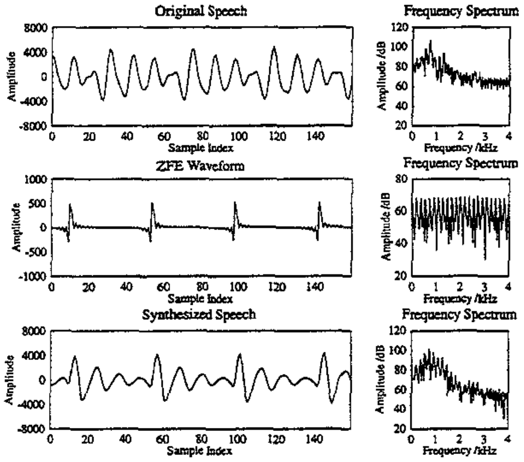

Figure 7: Time and frequency domain comparison of the original speech, ZFE waveform, and synthesized speech. The waveform is from an American female speaker.

the knowledge of λ, since the absolute location of the Zinc-pulses is irrelevant, as long as they are regularly spaced between the consecutive prototype segments. Hence, in contrast to the codec proposed by Hiotakakos and Xydeas [3], in our codec the Zinc-pulse position λ is not transmitted. Similarly to the λ parameter, the absolute location of the pitch prototype segment is only important at the encoder, again, for the ZFE optimization process. The interpolation process at the decoder assumes that the prototype segments are 20ms apart, but never considers their absolute time-domain location. Having characterized the basic features of our codec, let us now consider its performance and bit allocation scheme in the next Section.

4. CODEC PERFORMANCE AND CONCLUSIONS

The speech segment displayed in Figure 7 was recorded for an American female speaker, creating time and frequency domain waveforms that are typically voiced. Following the passage of the speech frame through the codec, in the time domain the overall shape of the original waveform is more or less preserved. However, the decay rate of the time domain signal resonances is quicker in the synthesized speech compared to the original waveform, most notably as regards to the second and third resonances. It has been suggested that this type of decaying signal benefits from adaptive post filtering [9]. In the frequency domain the overall spectral match is good, preserving both the spectral envelope and the fine structure shape. Observing the ZFE waveform of the centre-trace, a flat excitation frequency domain envelope is produced, while its spectral fine-structure reflects the pitch-dependent needle-like behaviour. Informal listening tests showed that the reproduced speech contained only slight "buzziness", but it was less transparent than the original speech.

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

Table 2: Bit allocation for the speech codec.

The bit allocation for the ZFE coder is summarized in Table 2, where 18 bits are reserved for LSF vector-quantization, while a one-bit flag is used for the V/U classifier. For unvoiced speech the RMS parameter is scalar quantized with 5-bits, the *transition* offset requires a maximum of 3-bits to encode the voiced-unvoiced transition point in terms of the number of voiced speech cycles within unvoiced frames. For voiced speech the pitch can vary from 20 → 147 samples, thus requiring 7-bits for transmission. The ZFE amplitude parameters A and B are scalar quantized with 6-bits.

5. REFERENCES

[1] W.B.Kleijn, "Encoding speech using prototype waveforms," *IEEE Transactions on Speech and Audio Processing*, vol. 1, pp. 386–399, October 1993.

[2] W.B.Kleijn and J.Haagen, "A speech coder based on decomposition of characteristic waveforms," in *Proceedings of ICASSP 95*, pp. 508–511, 1995.

[3] D.J.Hiotakakos and C.S.Xydeas, "Low bit rate coding using an interpolated Zinc excitation model," in *Proceedings of the ICCS 94*, pp. 865–869, 1994.

[4] Y.Hiwasaki and K.Mano, "A new 2-kbit/s speech coder based on normalized pitch waveform," in *Proceedings of ICASSP 97*, pp. 1583–1586, 1997.

[5] CCITT, *Coding of speech at 8 kbit/s using Conjugate-Structure Algebraic CELP*, G.729 ed., December 1995.

[6] F.C.A.Brooks and L.Hanzo, "Wavelet-based pitch detection for low-rate speech coding," submitted to *Proceedings of ICASSP 98*

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

[8] K.Yaghmaie and A.M.Kondoz, "Multiband prototype wavefrom analysis synthesis for very low bit rate speech coding," in *Proceedings of ICASSP 97*, pp. 1571–1574, 1997.

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