A NOVEL 1.7/2.4 KB/S DCT BASED PROTOTYPE INTERPOLATION SPEECH CODING SYSTEM

Prof. C.S Xydeas and H. Gokhan Ilk

Speech Processing Research Laboratory, Electrical Engineering Division,
Manchester School of Engineering, University of Manchester,
Manchester M13 9PL, U.K.

ABSTRACT

In this paper a novel DCT prototype interpolation synthesis process is presented and used to model the input speech signal. The compression efficiency of the DCT when applied to prototype pitch segments, leads to 1.7/2.4 kb/s DCT-PIC systems which can deliver decoded speech of high communication quality.

1. INTRODUCTION

In recent years low bit rate (<4.0 kb/s) speech coding techniques have been the focal point of considerable research activity as part of the quest for minimum bit rate, high quality speech. The applications of these efficient speech coding algorithms are numerous, ranging from voice mail to mobile voice and multimedia communication systems. In this context two main methodologies have emerged; ***i)*** Sinusoidal (SC) [1] and ***ii)*** Prototype Interpolation Coders (PIC) [2,3]. Although these methods, share certain speech modelling/synthesis principles, they lead into different codec structures.

Spectral modelling can be applied on the input speech directly or on the residual signal that is produced after removing, via an LPC analysis, short-term speech correlations. In the latter case the resulting LPC based vocoding system can deliver communication quality speech at 2.4 kb/s [2,3,4]. In this paper, a novel DCT prototype interpolation coding technique (DCT-PIC) is presented and applied directly on the input speech signal. The resulting coding system operates on speech frames of 20 ms duration and for each voiced frame a prototype waveform is selected. This provides a pitch cycle "representative" for the frame, which is then DCT transformed. Thus voiced frames are synthesised via a DCT based speech synthesis process that employs the DCT coefficients and the pitch estimates derived from adjacent frames. Furthermore, the system recovers accurately these "anchor" prototype segments and then synthesises the in-between waveform in a way that preserves perceptually important signal characteristics. Unvoiced frames are modelled via a LPC spectral envelope and an appropriately weighted random excitation sequence. Section 2 of the paper describes the proposed DCT-PIC in some detail, whereas section 3 addresses certain important issues which emerge from the quantisation of the codec parameters. The final section comments on system performance, complexity characteristics and also provides concluding remarks.

2. DCT-PIC SYSTEM DESCRIPTION

Figure 1 shows the overall encoder/decoder structure of the DCT-PIC system. Speech, sampled at 8 KHz, is divided in consecutive frames of size M=160 and a voiced/unvoiced decision process is carried out for each frame.

2.1 Unvoiced Speech

Unvoiced speech frames are recovered at the output of an 10^{th} order LPC synthesis filter H(z), whose input $A.G(i)$ is a normalised and scaled by A gaussian noise signal $G(i)$ .The value of A follows the energy of the short term LPC residual signal *Res(i)*. Thus $G(i) = G(i) \Big/ \sqrt{\sum_{i=1}^{N} G^2(i)}$, $G(i) \in Gaussian(\mu = 0, \sigma^2 = 1)$ and $A = \sqrt{\sum_{i=1}^{N} \mathrm{Re}s^2(i)}$. The scaling factor A is optimally quantized to $\hat{A}$ and then transmitted together with the corresponding quantized $L\hat{S}P_k$ $k = 1,2,...,10$ LPC filter coefficients. When $N<M$, the energy of the recovered unvoiced signal is adjusted more frequently than once, within a 20 ms interval.

2.2 Voiced Speech-Encoding

During the encoding of the current, say the nth, voiced frame, the encoder determines a "pitch prototype segment" $\{PS_n(m)\}$ for the frame. In general, the location and length P_n of this segment are transmitted to the receiver together with the quantized DCT coefficients $\{\hat{S}_n(k)\}$ $k = 0,1,...,P_n - 1$ of the prototype segment. At the receiver, an Inverse DCT based Interpolation process synthesises the recovered voiced signal over a period of T samples. T starts at the beginning of the (n-1)th frame prototype segment (i.e. the(n-1)th prototype) and ends at the sample prior to the beginning of the nth prototype, see Figure 2.

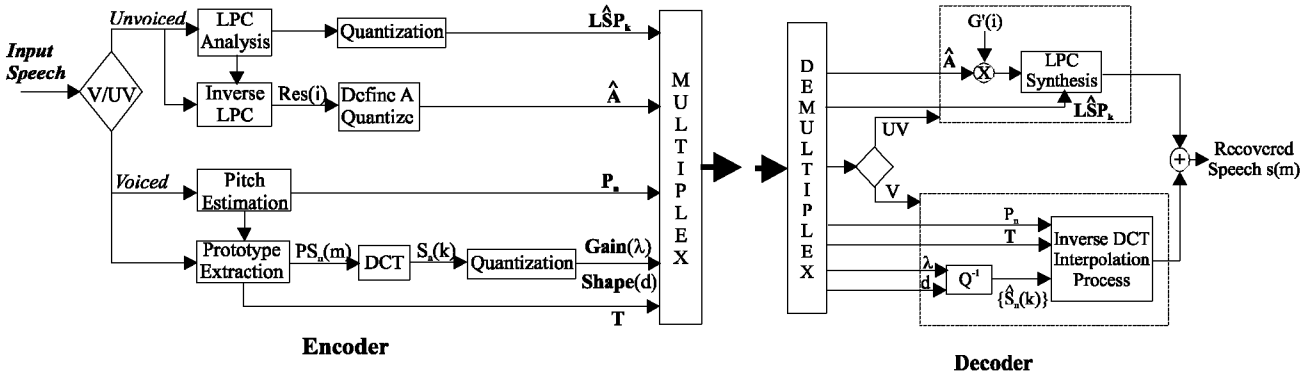

Figure 1 DCT-PIC Encoder and Decoder Description

Furthermore, a cross-correlation based process is employed in the selection of the nth prototype. This ensures alignment between successive prototypes and hence maximum cross correlation, a necessary condition for the efficient differential coding of prototype DCT coefficients. Thus a cross-correlation function $CR(\tau)$ is obtained between the (n-1)th prototype $\{PS_{n-1}(m)\}$ and a segment that commences at the beginning of the (n-1)th prototype and ends at the P_{max} sample following the end of the nth frame. P_{max} is the maximum expected pitch period in the input speech signal. $CR(\tau)$ is normalised with $CR(0)$ to $CR(\tau)$ and a peak $P(\tau)$ is selected in $CR(\tau)$, whose value is greater than 0.7 and whose location τ defines an interpolation interval T that is nearest to 20 ms. The nth prototype segment $\{PS_n(m)\}$ $m = 1,2,...,P_n$ is then defined by taking P_n speech samples starting from τ. Notice that system performance depends, to a large extend, on the accuracy of the pitch estimation process used to produce P_n. Thus the pitch estimation procedure employed in the system provides multiple estimates i.e. one every 5 ms and the nearest to the τ location is selected as P_n. A description of the pitch estimation process is beyond the scope of this paper.

2.3 Voiced Speech-Decoding

As mentioned in the previous section an Inverse DCT based Interpolation process is employed at the decoder to synthesise voiced speech $\{s(m)\}$ over an interval of T samples. This process is defined as

$$s(m) = \sum_{k=0}^{K(m)-1} f_k(u)^k \tilde{S}(k) Cos\left[\frac{(2a\ m+1)k\pi}{2P(m)}\right] \quad m = 0,1,...,T-1 \qquad \textbf{(1)}$$

$P(m)$ is a non-linear estimate of the pitch period at time m, $f_k = 1/\sqrt{2}$, for $k = 0$ and $f_k = 1$, for $k \neq 0$ whereas u=-1 when m belongs to even pitch segments and u=1 when m belongs to odd pitch segments. The sign inversion obtained for alternate DCT coefficients provides, during even pitch cycles, a "time domain" inversion of the signal which will be otherwise produced by a conventional DCT synthesis equation. The $\tilde{S}(k)$ function is obtained via the linear interpolation of the DCT sets of coefficients $\{\hat{S}_{n-1}(k)\}$ and $\{\hat{S}_n(k)\}$, i.e.

$$\tilde{S}(k) = \hat{S}_{n-1}(k)\left(1-\frac{m}{T}\right) + \hat{S}_n(k)\left(\frac{m}{T}\right) \quad m = 0,1,...,T-1$$

Notice that the number $K(m)$ of cosine terms and coefficients which contribute to the summation in Equation 1, is defined at each sampling instant as the nearest integer of the instantaneous pitch period $2\pi/\omega_2(m)$, for which the fundamental frequency $\omega_2(m)$ is the first derivative $d\phi_2(m)/dm$ of the phase $\phi_2(m) = \left(\frac{a\,k\pi}{P(m)}m\right)$. Thus aliasing effects are avoided since only terms which are within the Nyquist frequency are employed by the synthesis process. Furthermore, $P(m)$ is modelled by a second order interpolation function, in order to ensure that the following boundary conditions are satisfied,

1) The phase $\phi_k(m)$ of the cosine terms varies from 0 to 2π, over the length of a pitch period i.e.

$$Cos\left[\frac{(2a\ m+1)k\pi}{2P(m)}\right] \approx Cos\left[\frac{a'k\pi}{P(m)}m + \frac{k\pi}{2P_{n-1}}\right] = Cos\left(\phi_k(m) + k\theta\right)$$

and $\phi_k(0) = 0$, $\phi_k(T-1) = k\pi c_0$ where c_0 is the integer number of pitch cycles within T and

2) The fundamental frequency $\omega_2(m) = d\phi_2(m)/dm$ is equal to $2\pi/P_{n-1}$ at $m=0$ and $2\pi/P_n$ at $m=T-1$.

$$P(m) = \left(\frac{a}{T^2} - \frac{1}{c_0 T}\right)m^2 + \left(\frac{P_n}{T} + \frac{1}{c_0} - \frac{2a}{T}\right)m + a \qquad \textbf{(2)}$$

where $a = \frac{P_n P_{n-1} c_0}{T}$, and

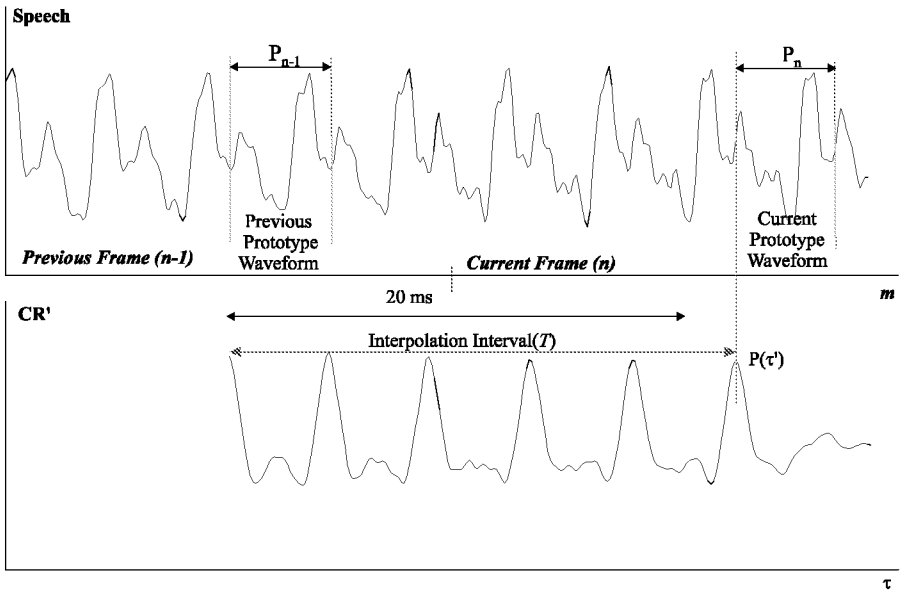

Figure 2 (n-1)th and nth Speech Frames and Cross-Correlation Function $CR(\tau)$

$$a = \frac{P_n c_0}{T} \tag{3}$$

The inclusion of a in Equation 1 allows for the time expansion ($a<1$) or compression ($a>1$) that is required when $P_n=P_{n-1}$ and T is not an exact multiple of P_n. Notice that the transmission of T to the receiver can be avoided, with a subsequent reduction in the bit rate of the system, by modelling $P(m)$ with a linear interpolation function. In this case

$$T = c_0\sqrt{P_n P_{n-1}} \tag{4}$$

and c_0 is the integer which ensures that T is as close as possible to 20 ms. Furthermore, Equation 4 is also used to define the interpolation interval T during transitions from voiced to unvoiced frames and vice versa. In this case $T=c_0P$ where P is the pitch period of the voiced frame. Notice that during these transitions there is an "overlap" between segments produced by ***i)*** the "voiced speech" DCT synthesis process and ***ii)*** the "unvoiced speech" LPC synthesis process. In this case the two linearly weighted components are combined by an overlap-add process that operates over the overlap period. The prototype DCT coefficients used for the unvoiced frame are all set to zero.

3. QUANTIZATION OF PROTOTYPES

The quantization of sets of prototype DCT coefficients, each set having its own dimension P_n, is an important coding task. Since prototype signals are compressed efficiently by the DCT process, the values of higher frequency coefficients are relatively small, and can be therefore quantized less accurately than lower frequency coefficients. To this end, it has been verified experimentally that high communications quality speech is obtained from the system while coding accurately the first 30% of the available DCT coefficients. The remaining coefficients can be represented by a single value which ensures that the energy of the synthesised speech signal follows the energy of the input speech.

Of course, synthesised speech quality improves when coding accurately all DCT coefficients. The best way to achieve this is via differential coding techniques since, due to the alignment process employed in the system, successive prototypes are similar.

In this case an error is formed as

$$E_n(k) = S_n(k) - a_n\hat{S}_{n-1}(k) \qquad k = 1,2,\ldots,P_n - 1 \tag{5}$$

Here the linear prediction coefficient a_n is defined as

$$a_n = \frac{\sum_{j=1}^{P_n-1} S_n(j)\hat{S}_{n-1}\left\lfloor j\frac{P_{n-1}-1}{P_n-1}\right\rfloor}{\sum_{j=1}^{P_n-1} \hat{S}^2_{n-1}\left\lfloor j\frac{P_{n-1}-1}{P_n-1}\right\rfloor} \tag{6}$$

where $\lfloor\ \rfloor$ denotes a nearest lower integer operation and effectively allows for different sizes of sets of coefficients. Alternatively a fixed coefficient α can be used in Equation 6 whose value has been determined experimentally as α=0.8. The resulting error sequence $E_n(k)$ $k=1,2,\ldots,P_n-1$ is then quantized using a Variable Dimension [5] Shape (d), Gain (λ) Split Vector Quantizer. Assuming that the maximum expected pitch period is P_{max}=120, the sizes of the vectors in the fixed codebooks, with which parts of the variable size input vectors are compared, are set to C_1, C_2, ..., C_b. b indicates the number of "bands" used to split the codebook elements and $C=C_1+C_2+\ldots+C_b$, $C\geq P_{max}$, and

$C_1 \le C_2 \le \ldots \le C_b$. The λ_i i=1,2,...,b gains produced by the process are log converted and vector quantized.

At the receiver the decoded $\hat{\lambda}_i$ and $\hat{d}_i$ values are used to recover $\{\hat{S}_n(k)\}$ which then drives the Inverse DCT Interpolation process.

Notice that the number of bits Cb_i i=1,2,...,b allocated to each "band" can be fixed to a value which, given a codec bit rate, can be defined experimentally for maximum overall system performance. Notice however that as the codec bit rate decreases, the number of bits Cb_i allocated to a band might not be sufficient to yield a minimum required quantization accuracy for that band. Furthermore quantization accuracy in this case depends on the size P_n of the input vector. Thus whereas in the case of an input vector with below average pitch period size, the number of error $E_n(k)$ samples "es_j" which "fall" in the i^{th} band might be low enough to enable an adequate representation with Cb_i bits, an above average size input vector will yield a larger number of error samples which can be quantized less accurately while using the same number of bits, i.e. Cb_i. This dependency of quantization accuracy to the input vector size emerges because of the phase information, that is conveyed by the DCT coefficients, and the relative limitations of the prototype alignment process. In turn, this indicates that the range P_{min} to P_{max} of input vector sizes can be divided into bands Pb_i. The system can then employ different vector $\{C_i\}$ split and $\{Cb_i\}$ bit allocation strategies for each Pb_i range. This provides an extra degree of adaptivity in the prototype quantization process at the expense of an increase mainly in storage requirements.

4. RESULTS AND CONCLUSIONS

DCT-PIC results obtained from computer simulations clearly indicated the potential of the system is providing high communications quality speech in the range of 1.7 to 2.4 kb/s. At 1.7 kb/s while adopting the bit allocation strategy of Table 1, the system uses a fixed prediction coefficient α in producing $E_n(k)$, C=100, C_1=15, C_2=15, C_3=30 and C_4=40. 9 and 8 bit codebooks are used to quantize the first two parts of the error vector $E_n(k)$. The next two parts are represented by the corresponding RMS values of the DCT coefficients of these parts. Thus λ_3=RMS_3 and λ_4=RMS_4. At the decoder the last two vector parts are recovered as $\hat{E}_n(k) = \hat{\lambda}_3 G(k)$ and $\hat{E}_n(k) = \hat{\lambda}_4 G(k)$. Also 9 bits are allocated for the quantization of these four λ_i gains. The bit allocation of the 2.4 kb/s system includes 3 bits for the transmission of a_n, and also two extra codebooks of size 6 and 5 bits for coding the last two parts of $E_n(k)$.

Notice that, as expected, the accuracy with which the $S_n(k)$ coefficients are quantized affects the "envelope" of the magnitude spectrum of the reconstructed speech signal. The bit allocation strategy of Table 1 is by no means optimised for maximum coding performance. Work is currently under way in order to define optimal fixed as well as adaptive (multiple split VQ scenarios depending on P_n) prototype quantization procedures.

Alternatively, The DCT-PIC synthesis process can be used on the LPC residual signal. In this case LPC "envelope" information can be transmitted with 800 to 1000 bits while the remaining bits represent the "fine" structure of the speech short term magnitude spectrum.

	Unvoiced	Voiced (1.7 Kb/s)	Voiced (2.4 Kb/s)
Pitch	-	7	7
V/Unvoiced Flag	1	1	1
Adaptive Predictor	-	-	3
Gain	7	9	9
Shape 1,2,3,4	-	9,8,0,0	9,8,6,5
LSF's	26	-	-
Total Bits/20msec	34	34	48

Table 1 DCT-PIC Bit Allocation Table

REFERENCES

[1] **R.J. McAulay and T.F. Quatieri,** "*Low-Rate Speech Coding Based on the Sinusoidal Model*", Advances in Speech Signal Processing edited by S.Furui and M.M. Sondhi, Marcel Dekker Inc., pp. 165-208.

[2] **W.B. Kleijn and J. Haagen,** "*A Speech Coder Based on Decomposition of Characteristic Waveforms*", IEEE, ICASSP-95, page 508.

[3] **C. Papanastasiou and C.S. Xydeas,** "*Efficient Mixed Excitation Models in LPC Based Prototype Interpolation Speech Coders*", IEEE, ICASSP-97 Vol.2, page 1555.

[4] **A. McCree, K. Truong, E.B. George, T.P. Barnwell and V. Viswanathan,** "*A 2.4 Kbit/s MELP Coder Candidate for the New U.S. Federal Standard*", IEEE, ICASSP-96, page 200.

[5] **A. Das and A. Gersho,** "*Variable Dimension Vector Quantisation of Speech Spectra for Low-Rate Vocoders*", IEEE Proceedings of the Data Compression Conference, pp. 420-429, April 1994.