Proc. Biennial Symp. Commun. (Kingston, ON), pp. B5.9–B5.12, June 1984

Computational Considerations in Adaptive Transform Coding of Speech

PETER KABAL and RAFI RABIPOUR

In recent years increasing emphasis is being placed on the digital encoding of speech and other analog signals. Adaptive transform coding (ATC) of speech signals offers high quality signal reproduction at low to moderate transmission rates (8–16 kb/s) [1]. In this paper, computational aspects of ATC are examined with the goal of developing an algorithm which produces high speech quality and can be implemented in practical real-time hardware.

In transform coding, the input speech samples are processed in blocks (frames). Each block of samples undergoes a linear transformation. The transformed coefficients are quantized one at a time and coded for transmission. At the receiver, the received data is decoded and passed through an inverse transformation. The resulting samples represent the coded speech signal. The transformation used in ATC is the discrete cosine transform (DCT). It has the property of approximately decorrelating a highly correlated first order Markov sequence. The DCT can be computed efficiently using algorithms based on the fast Fourier transform (FFT) algorithm.

Computational Steps in ATC

a) Preemphasized speech samples are grouped into overlapping frames of from 32 to 256 samples for speech sampled at 8 kHz. The DCT is applied to this frame of data.

b) The transform coefficients are quantized one by one. The number of bits used to quantize each coefficient is determined by the bit assignment procedure so as to result in the best quality reproduced speech.

c) The quantized coefficients, along with the side information necessary to determine the bit assignment, are transmitted to the decoder.

d) The decoder computes the bit assignment based on the side information using the identical algorithm to that used in the transmitter.

e) Using the bit assignment, the decoder parses the incoming bit stream for the coded DCT coefficients. The decoder reconstructs the transform coefficients from the decoded information.

f) The inverse DCT is taken of the transform coefficients. The resultant data is deemphasized (an inverse operation to preemphasis) and windowed to give the reconstructed speech.

1. Spectral Modelling

The major influence on the quality of the reconstructed speech is the bit assignment strategy. A smooth spectral fit to the DCT coefficients provides an estimate of the energy of the coefficients. This energy estimate serves a dual role. First, it is used to determine the number of bits to be assigned to each coefficient in such a way as to minimize the total mean-square quantization error. Second, the spectral estimate is used to scale the quantizer, in effect adapting the quantizer range to the coefficient being quantized.

The envelope of the DCT spectrum retains the formant structure of speech. The problem of modelling the spectrum is similar to one which occurs in linear predictive coding (LPC) of speech.

This work was supported in part by university contracts from the Communications Research Center, Department of Communications.

P. Kabal is with the Department of Electrical Engineering, McGill University, Montreal, Quebec, H3A 3A4 and INRS-Télécommunications, Université du Québec, Verdun, Quebec, H3E 1H6.

R. Rabipour was with INRS-Télécommunications, Université du Québec. He is now with BNR, Verdun, Quebec, H3E 1H6.

An all-pole model is used since it tends to highlight the spectral features of the speech signal known to be perceptually important. The spectral model is determined most conveniently from autocorrelation coefficients. In this case, a pseudo-autocorrelation function is determined by taking the inverse Fourier transform of the DCT energy spectrum. The optimal predictor filter coefficients are determined from the pseudo-autocorrelation terms using the Levinson recursion algorithm. This linear prediction modelling minimizes the error term

$$\epsilon^2 = \sum_{n=0}^{2N-1} \left[y_n - \sum_{i=1}^{M} a_i y_{(n-i)} \right]^2 , \tag{1}$$

where M is the order of the predictor filter. For the present case, it is the all-pole LPC synthesis filter (the inverse of the predictor filter) which models the DCT spectrum. The basis spectrum is the energy spectrum of the synthesis filter,

$$|H_k|^2 = \frac{g^2}{\left| 1 - \sum_{i=1}^{M} a_i e^{-j2\pi ik/N} \right|^2} , \tag{2}$$

where g is the filter gain. For computational purposes, this spectrum can be calculated as the term by term inverse of the square of the DFT of the predictor filter response.

1.1 Basis Spectrum Weighting

The error term which is minimized by the LPC modelling can be expressed in the frequency domain as the sum of the ratios of the actual energy spectrum, X_k^2, to that of the basis spectrum, $|H_k|^2$,

$$\epsilon^2 = \frac{g^2}{N} \sum_{k=0}^{N-1} \frac{X_k^2}{|H_k|^2} . \tag{3}$$

The mean-square coding error can be frequency-weighted to take advantage of the properties of human auditory perception. Since more noise can be tolerated in the formant regions than in the valleys between formants, a weighting which deemphasizes the formant regions is chosen. The weighting is implemented by raising the basis spectrum to the power α and using this weighted basis spectrum in the bit assignment computation. $\alpha = 0$ results in a constant bit assignment. $\alpha = 1$ results in a distribution of bits that minimizes the unweighted mean-square error, i.e. the distortion is approximately constant with frequency. A compromise value produces the "best" sounding speech.

1.2 Pitch Modelling

The spectral distribution of the assigned bits can be improved for speech signals by taking into account the quasi-periodicity of the speech signal at the pitch rate during voiced segments. The pitch structure manifests itself in the frequency domain as a superimposed comb spectrum. The scheme chosen to incorporate pitch information is a combined formant/pitch model. This method uses the low-order correlation terms (as in the usual LPC fit) but also uses the correlation terms centred around the pitch peak. A combined set of formant/pitch terms is formed to simultaneously solve for the spectral fit.

1.3 DCT Coefficient Bit Assignment

The bit assignment is determined so as to minimize the mean-square error for a frame of samples. A direct approach to assigning integer numbers of bits is used. The procedure assigns bits in such a way as to maximally decrease the mean-square error with each additional bit that is assigned. This procedure can also be used with empirically determined (tabulated) rate distortion functions.

2. Computational Considerations

The arithmetic in DSP hardware is usually implemented as *fixed point* computations. The fixed point arithmetic can result in *round-off* and/or *truncation* error. In the sequel we assume 16-bit accuracy.

2.1 The Discrete Cosine Transformation

A major concern in the fixed point computation of the DCT is the problem of register overflow. For the fixed point DCT of a complex sequence $\{ f_n \}$ of length N, register overflow is prevented if in the fractional notation $|f_n| < 1/N$. If $|f_n|$ itself is less than 1, an obvious way to prevent overflow is to scale the input data by $1/N$. However, this method results in a poor signal-to-noise ratio (SNR) due to the fact that $\log_2 N$ bits of precision have been discarded at the beginning of the computation. A better approach is to distribute the scaling over the various stages of the FFT computation.

In order to reduce the truncation error due to such scalings the input sequence is prescaled by a power of two to bring the fractional value of the largest component to between 0.25 and 0.5. For transform lengths of 256, noticeable noise is introduced in the overall coding and decoding system if extreme care is not exercised in the scaling of the DCT computation.

2.2 Pseudo-autocorrelation Function and Reflection Coefficients

The straightforward application of fixed point arithmetic to calculate the pseudo-autocorrelation function (used in the determination of the spectral fit) using an FFT results in serious deterioration of the output speech quality. The dynamic range involved cannot be represented with only 16 bits of precision. An alternative approach is to compute the pseudo-autocorrelation function directly from the input sequence. The pseudo-autocorrelation function can be obtained directly from the input data using

$$R_n = \frac{1}{2} \sum_{i=0}^{N-1-n} x_i x_{i+n} + \frac{1}{4} \sum_{i=0}^{n-1} \left[x_i x_{n-1-i} + x_{N-n+i} x_{N-1-i} \right] \qquad 0 \le n < N \,, \tag{4}$$

Since only the first few values of R_n are needed to solve for the reflection coefficients, the number of computations needed for a direct computation is modest—in fact smaller than for an FFT approach for reasonable parameter values. Perhaps the most important benefit to be gained from the direct method is a substantial increase in accuracy—the autocorrelation computation can take full advantage of the double-precision product and accumulate feature available on DSP chips. The main disadvantage of the direct method is that the calculation of the pitch period requires considerable extra effort. For this reason, pitch modelling is reluctantly abandoned for a fixed point implementation.

Problems remain for fixed point computations for frames with a large dynamic range. For such frames the correlation equations may be numerically ill-conditioned. This problem can be alleviated easily by adding low-level noise to the DCT spectrum. The effect of adding noise to the DCT spectrum of the input signal is achieved in the implementation by adding a small fraction of the first pseudo-autocorrelation term to itself.

The algorithm for the actual solution of the linear equations which determine the reflection coefficients uses a set of intermediate variables (corresponding to the normalized prediction error), that have magnitude less than unity [2]. This normalization is fully compatible with fractional fixed point representation of the data.

2.3 Computation of the Basis Spectrum

There are two major problems involved with the calculation of the basis spectrum from the filter coefficients in fixed point arithmetic using an FFT. The first is the loss of $\log_2 2N$ bits of precision due to the normalization required at each stage of the FFT. The other problem is the squaring operation required to calculate the magnitude of the response. The numerical errors have

the greatest impact on the regions of the spectrum where the magnitude of the predictor filter is small. Such regions are the high amplitude parts in the basis spectrum and hence are assigned most of the bits.

The method chosen to circumvent these problems calculates the inverse of the basis spectrum from the autocorrelation of the predictor filter. The predictor filter has only $M+1$ coefficients and so its correlation function is short and inexpensive to compute. The DFT of this correlation function can be written as follows for $M < N$,†

$$\frac{1}{|H_k|^2} = R_{aa}(0) + 2\sum_{n=1}^{M} R_{aa}(n)\cos\frac{\pi nk}{N} \qquad 0 \le k < N\ . \tag{5}$$

In addition, the direct formulation lends itself to efficient and accurate computation with a DSP chip since the double-precision product and accumulate feature of the DSP chip can be used.

Even as described, this method does not yield sufficiently accurate results. A solution is to use mixed precision computations—the $R_{aa}(.)$ values are represented as $M+1$ double-precision values while the cosine terms are single precision 16-bit scaled integer values. The computation is arranged so that the accumulated sum of the products of the cosine terms and the low-order part of the correlation values is shifted by 16 bits and added to the accumulated sum of the products of the cosine terms and the high-order part of the correlation values. In this way, mixed precision computations require approximately twice as many operations as would be required normally, but less than full double-precision computations.

The double-precision results are converted to 16-bit values by extracting the most significant part of the result shifted to the left so as to represent the smallest (hence the most important) values with at least two bits. The samples corresponding to larger components which overflow the 16-bit registers as the result of the left shift operation are set to the largest positive 16-bit value.

3. Number of Operations

The number of operations needed for an ATC coder is tallied in Table 1 for several cases. These counts are for the basic arithmetic operations involved in the signal processing. They do not include scaling operations which are numerous and essential to the correct operation of a fixed pont algorithm. For each case in the table, the spectral estimates used for bit assignment are updated every 256 samples. The subframe size indicates the number of samples in the transform. For a subframe size of 64, four subframes are averaged for the spectral estimate. These operations counts indicate that an ATC coder can be implemented with a relatively simple hardware configuration using commercially available DSP chips.

Subframe size	No. Bits /frame	Coder No. Multiplies	Coder No. Additions	Decoder No. Multiplies	Decoder No. Additions
256	244	406K	2463K	337K	2393K
128	122	315K	843K	244K	773K
64	61	265K	411K	194K	340K

Table 1 Number of Operations (Thousands) per Second

References

1. J. M. Tribolet and R. E. Crochiere, "Frequency domain coding of speech", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 512–530, Oct. 1979.

2. J. LeRoux and C. J. Gueguen, "A fixed point computation of partial correlation coefficients", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 257–259, June 1977.

† A $2N$-point DFT is required to obtain the N-point basis spectrum.