ADAPTIVE TRANSFORM CODING FOR LINEAR PREDICTIVE RESIDUAL

D. J. Mudugamuwa and A. B. Bradley
Department of Communication and Electronic Engineering
RMIT, PO Box 2476 V, Melbourne 3001, Australia.
Tel: 61 3 96602455, FAX: 61 3 96621060
E-mail:damith@catt.rmit.edu.au, alanb@rmit.edu.au

ABSTRACT

In voice coding applications where there is no constraint on the encoding delay, such as store and forward message systems or voice storage, segment coding techniques can be used to achieve a reduction in data rate without compromising the level of distortion. For low data rate linear predictive coding schemes, increasing the encoding delay allows one to exploit any long term temporal stationarities on an interframe basis, thus reducing the transmission bandwidth or storage needs of the speech signal. Transform coding has previously been applied to exploit both the inter and intra frame correlation, but has been limited to short term spectral redundancy [1][2]. This paper investigates the potential for data rate reduction through extending the use of segment coding techniques to identify redundancies in the LPC residual. Initial tests indicate a potential 40% average reduction in data rate for a given subjective speech quality.

1. INTRODUCTION

Due to the non-stationary behaviour of speech, a linear analysis/synthesis model can only be employed accurately over a small time period, generally in the range 10 - 35 ms. During this period, model parameters must be updated at least once. During certain phonetic combinations however, the speech signal can exhibit a greater degree of stationarity extending over a period of up to several hundreds of milliseconds. Consequently during these periods, there is significant correlation between successive frames of the model parameters and it is possible to exploit this correlation to reduce the overall bit rate at the expense of added coding delay.

Linear Predictive (LP) based speech coding algorithms transmit two information components, a short time spectrum estimate and an excitation or residual signal. For all experimental and testing work the Mixed Excitation LPC (MELP) vocoder described in [3] and [4] has been employed. In the MELP coder the short term spectral information is represented by line spectral pair (LSP) frequencies and excitation is generated by mixing noise and a harmonic signal in different proportions in five non overlapping frequency bands. The proportions in each band are evaluated by measuring the voicing strength of the speech signal in the particular frequency band. The frame duration has been kept to 22.5 ms in this experimental work. In the MELP coder of [4], the encoded parameters transmitted for each frame are represented as 6 separate vector quantities as follows;

a) 10 LSP's, carrying the short term spectral information,
b) 5 voicing strengths, carrying the noise/harmonic ratio in each of five bands,
c) one pitch value, describing the period of the harmonic signal,
d) 10 Fourier coefficients carrying additional spectral information to account for the imperfections of the LP modelling,
e) 2 energies, carrying pitch synchronous energy per sample per each half frame and
f) a binary decision indicating the onset of a jittery voicing state.

The block diagram of the simulation setup is shown in figure 1. Speech data is read from the TIMIT database, lowpass filtered at 3.4 kHz, decimated to 8 kHz and fed to the MELP coder. The MELP coder generates the 6 model vectors described above and these are buffered to a depth of 20 frames (450 ms). For simulation purposes the encoding and decoding sequence is integrated and processing is focussed on a single data buffer as indicated in figure 1.

2. SEGMENTATION

The buffered frames of vector parameters are segmented by identifying the boundaries of voiced, unvoiced and silence regions of the speech signal. The voiced-unvoiced decision is made similar to LPC10e by considering the Average Magnitude Difference Function (AMDF) windowed maximum-to-minimum ratio, the zero crossing rate, energy measures, prediction gains and

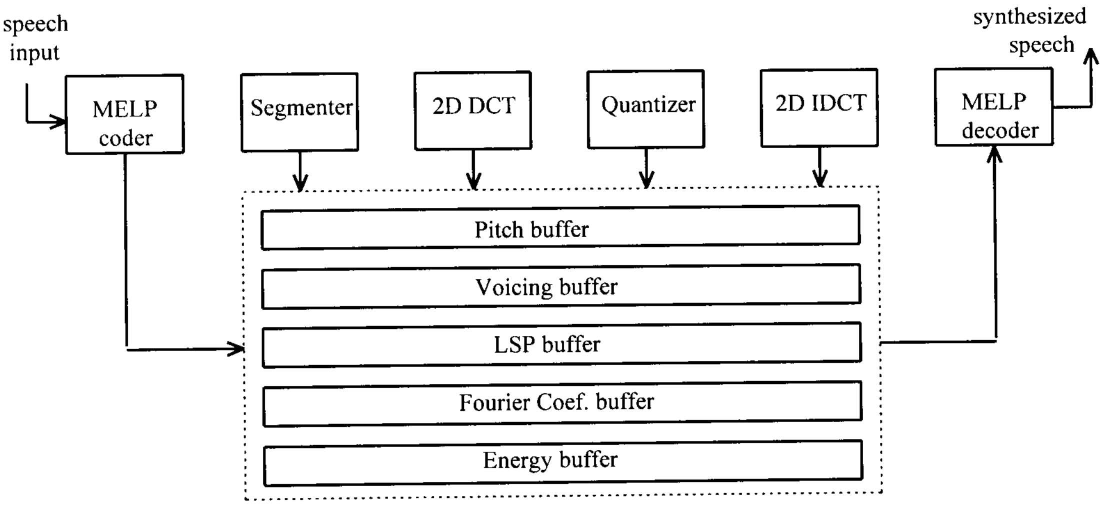


FIGURE - 1 Segmentation / Quantisation simulation setup

reflection coefficients. Silence classification is based on a comparison of the current frame energy with an adaptive threshold determined over the previous 500 frames. Classifying the segmentation process in this way allows the use of different bit allocations for different segment types and also permits efficient quantisation by normalising the parameters using their individual mean and variance values evaluated for different segment types. The maximum segment sizes are limited to 20 frames for silence and voiced speech and 8 frames for unvoiced speech. Segmenting speech into blocks of like frames provides a two dimensional data source appropriate for adaptive transform compaction.

3. PARAMETER VECTOR TRANSFORMATION

For each vector parameter segment a two dimensional discrete cosine transform (2D-DCT) is applied over the segment length. One dimension provides for the successive frames of a segment whilst the second dimension contains the elements of the parameter vector within the frames. This allows exploitation of both inter frame and intra frame correlation amongst the different parameter elements to achieve a data compaction. For the pitch parameter only a 1D-DCT is utilised.

The binary jittery voicing state and the segment type information are not subjected to transform coding.

Half frame pitch synchronous energy estimations are converted to a logarithmic scale before the transform is applied and this results in more manageable transform coefficients for the quantization. Bandpass voicing estimates are not rounded off to 0 and 1 as in [4] but allowed to take continuous values between 0 and 1, resulting in a more appropriate voicing data vector for transform coding.

4. QUANTIZATION

De-correlated transformed coefficients are normalised to zero mean and unit variance, and scaler quantized. Mean and variance for each transformed coefficient for different segment sizes and types are predetermined by a training process using 227, randomly chosen TIMIT training speech files to include 8 dialect regions, and available at the encoder and the decoder. Segmentation allows the parameters to be normalised using individually determined mean and variance for voiced, unvoiced and silence cases, optimising the quantisation.

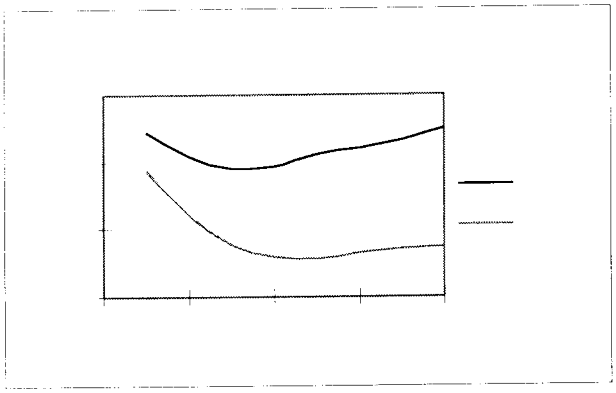
A Lloyd-Max quantizer has been designed [5][6] using the probability density functions (pdf) obtained from the transformed coefficients themselves. The distribution of each of the normalised transform coefficients within a particular parameter was found to have a similar shape regardless of position in the data block. Further, these distributions are also observed to be symmetrical. These two facts reduce the quantizer design and storage complexity dramatically. However for any chosen bit rate and thus every unique bit allocation for every parameter, a separate quantizer design must be available.

Quantizer overloading has a large impact on the synthesized speech quality. Since the optimal quantizer loading factor depends on both the pdf distribution shape for the particular parameter coefficients and the chosen bit rate of the quantizer it is necessary to determine the optimal loading factor for each quantizer separately. A suboptimal procedure is adopted to determine the loading factors as follows. For each parameter an optimal loading factor is determined for two different bit rates and the value is interpolated linearly for other bit rates. Spectral distortion (SD) given by [8],

$$- \Big| \qquad \Big| \qquad ---(1)$$

is calculated and averaged over successive frames, for different loading factors. *Y(.)* and *X(.)* are quantized and unquantized versions of the synthesized speech spectrum.

The loading factor that minimises the averaged SD is selected as the optimal value for the bit rate considered. Selection of loading factor is a trade off between granular noise and overload distortion. Figure 2 shows variation of average SD against the loading factor for the LSP transform coefficients for two bit rates, 2.2 and 4.8 bits per coefficient.



5. BIT ALLOCATION

For each transform coefficient, bit allocation is determined by it's variance, according to [7] and these bit allocation concepts are extended to multiple segment sizes within a particular segment type and a particular parameter. The bit allocation for the i^{th} transform coefficient in j^{th} frame of a block of NxM (ie. block of M frames of a parameter having a dimension N) is given by,

$$- \quad \overline{\qquad} \qquad ---(2)$$

where is the variance of the transform coefficient and b_1, GM_1 are the average bit allocation per coefficient and the geometric mean of transform coefficient variances of the single frame block (ie. M=1 case) for the particular parameter. These bit allocations ensure the same quantizer noise variance amongst all the transform coefficients of a parameter. When the average bit rate for a single frame block, b_1, is specified for a parameter, then the bit allocations for every transform coefficient of the particular parameter can be calculated for every possible block size.

These bit allocations are optimal in a mean square sense for each available block size, enabling lower average bit rates for larger block sizes due to the transform coding gain.

For the silence segments only the energy parameter is needed to be quantized.

For a target composite data rate the proportioning of allocated bits to the various parameters, in the single frame block case, ie. the determination of b_1 for each individual parameter and segment type, was optimised for best subjective quality and results are shown in table 1.

The actual bit rate achieved when segments of multiple frames are allowed will be lower than that suggested by the figures of table 1, which are for the single frame per segment case.

Table 1 Average bit allocation per transform coefficient for single frame segment parameters

Parameter	Unvoiced	Voiced	Silence
Pitch	4	4.5	0
Voicing	4	4	0
LSP's	2.5	2.5	0
Fourier coef.	0	0.8	0
Energy	4.5	4.5	2

The synthesis process decodes the quantized transform coefficients according to stored reconstruction values and denormalises using stored mean and variance for each possible transform coefficient. Parameters assigned zero bits are reconstructed from stored mean value along.

6. PARAMETER SCALING

The bit allocation strategy defined by equation (2) minimises mean square quantization error in the transform domain, ensuring uniform distribution of quantization error variance across each element of the original parameter vector. This can be a disadvantage when control of quantization resolution across the various elements of a parameter could yield a perceptual enhancement of recovered speech quality.

This difficulty can be solved by scaling the parameter elements before the transformation. This process effectively changes the variance of the parameter element before the transform is applied. An inverse scaling should be carried out after the inverse transform. If a particular parameter vector element is scaled up before the DCT is taken, it is quantized with a finer resolution than other elements of that parameter. This requires the encoder and the decoder to store additional scaling factors as side information.

To investigate the potential benefit of this concept low frequency element values of the LSP's, voicing and Fourier coefficient parameters were scaled upwards prior to transformation to enhance relative resolution. Subjective results for this test did not however provide significant improvement to recovered speech quality and this strategy has not been included in the final implementation.

7. SIMULATION RESULTS

Subjective and objective tests have been carried out using TIMIT database test speech files selected from 8 dialect regions.

For the evaluation of the proposed coding scheme a scaler quantized version of the MELP coder, that does not exploit the inter frame and intra frame redundancies of the parameters, is also implemented. The final bit rates for the two coders; DCT and scalar quantized versions, are selected so that the majority of listeners cannot differentiate the quantized and unquantized versions of the synthesized speech output. Results of the subjective tests have shown a need for 72 bits per frame for the scaler quantized case and 35 bits per frame for the DCT case to satisfy this criteria. The observed actual average number of bits used per frame for the proposed DCT coding scheme are shown in table 2 for continuous speech.

TABLE 2 Actual average bit use per frame

Parameter	Bits/frame
Pitch	3.05
Voicing	2.78
LSP's	17.57
Fourier coef.	1.85
Energy	6.98
Jittery state	1
Total	33.23

In addition to the parameters in table 2, it is also required to quantize the segmental information. Segmental overhead is 6 bits per segment and this represents an approximate overhead of 1.25 bits/frame. Since the frame duration is 22.5 ms, this represents an overall data rate of 1533 bits/second.

8. CONCLUSION

In this paper a transform coding scheme has been presented to exploit the delay domain in quantizing the LP residual. The proposed scheme extends earlier work[1] using transform coding of LP residual and has been implemented on a MELP platform. Informal subjective testing has confirmed 50% reduction in transmission bit rate over the scaler quantized case for continuous speech with natural pauses.

Further improvements can be made to this scheme in the areas of segmentation, pitch quantization and jittery state quantization.

9. REFERENCES

[1] Glazebrook, E., Bradley, A.B., "Low data rate adaptive transform coding for parametric representation of speech signals", *ISSPA 1996.*

[2] Farvardin, N., Laroia, R., 1988, "Efficient Encoding of Speech LSP Parameters Using the Discrete Cosine Transform", *IEEE Transactions on Speech and Audio Processing*, 1989, pp.168-171.

[3] McCree, A.V., Barnwell III, T.P., 1995, "A mixed excitation LPC vocoder model for low bit rate speech coding", *IEEE Transactions on Speech and Audio Processing,* vol.3, no.4, pp242-249.

[4] McCree, A.V., Truong, K., Geoge, E.B., Barnwell III, T.P., Viswanathan, V., 1996, "A 2.4 BIT/S MELP coder candidate for the new U.S. Federal Standard", *International Conference on Acoustics, Speech and Signal Processing,* Atlanta, Georgia, May 7-9, 1996, pp.200-203.

[5] Max, J., 1960, 'Quantization for minimum distortion', *IRE Transactions on Information Theory,* pp.7-12.

[6] Lloyd, S.P., 1982, 'Least squares quantization in PCM', *IEEE Transactions on Information Theory,* pp.129-136.

[7] Jayant, N.S., Noll, P., 1984, " *Digital Coding of Waveforms*", Prentice-Hall, Signal Processing Series Prentice-Hall, Inc., New Jersey.

Svendsen, T., 1994, 'Segmental quantization of speech spectral information', *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1994, pp. I.517-I.520.