ADAPTIVE TRANSFORMATION FOR SEGMENTED PARAMETRIC SPEECH CODING

Damith J. Mudugamuwa, Alan B. Bradley
Dept. of Communication and Electronic Engineering
RMIT, Victoria 3001, Australia.

ABSTRACT

In voice coding applications where there is no constraint on the encoding delay, segment coding techniques can be used to achieve a reduction in data rate. 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 in low data rate speech coding to exploit both the interframe and the intraframe correlation [1][6][8]. This paper investigates the potential of an adaptive transformation scheme for a segmented parametric speech representation. The problem of transform quantization is formulated and a solution methodology was proposed. The potential benefit of the use of the proposed adaptive transformation scheme is discussed in the context of segmented LSPs.

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 msec. During certain sustained phoneme elements 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.

Segmentation techniques together with the Discrete Cosine Transform (DCT) were employed to quantize the MELP vocoder [5] parameters at 1530 bps, at the expense of 450 msec coding delay, so that at a 95% confidence limit, listeners could not differentiate the quantized and unquantized versions of the synthesized speech output for 85% of the test phases. Further improvement for this coding scheme was made by introducing the optimal Karhunen-Loeve Transform (KLT), in a fixed average sense in place of the DCT [6][7]. This paper discusses the development of an adaptive transformation scheme with locally optimal transforms in place of the non-adaptive globally optimal KLTs in [7].

2. TRANSFORM CODED MELP MODEL

The MELP model generates 6 parameter vectors per frame (22.5 msec). Namely 10 LSPs, 5 voicing strengths, 2 energies, 10 Fourier coefficients, a pitch value and a jittery voicing state. These parameters were buffered to a depth of 20 frames.

The buffered frames of vector parameters were segmented into blocks by identifying the boundaries of voiced, unvoiced and silence regions of the speech signal. The voiced-unvoiced decision was made similar to LPC10e and silence classification was based on a comparison of the current frame energy with an adaptive threshold determined over the previous 500 frame energies. The maximum block sizes were limited to 20 frames for silence and voiced and 8 frames for unvoiced speech. Segmentation was implemented in a way to ensure no fragmentation of the blocks occur due to the limited buffer size.

For each parameter block a two dimensional (2D) transformation was applied. One dimension provides for the successive frames of a block 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. The binary jittery voicing state and the block type information were not subjected to the transform operation. 2D transformation was implemented by applying two, one dimensional (1D) transforms row wise and column wise for the 2D parameter blocks.

De-correlated transformed coefficients were normalised to zero mean and unit variance, and scaler quantized. Mean and variance for each transformed coefficient for different block sizes and types were predetermined by a training process and available at the encoder and the decoder. Lloyd-Max quantizers [3][4] were designed using the probability density functions (pdf) obtained from the transform coefficients themselves.

For each transform coefficient within a parameter, bit allocation was determined by it's variance, according to [2][6] and are optimal in a mean square sense amongst all available block sizes, enabling lower average bit rates for larger block sizes due to the transform coding gain. For the silence blocks, only the energy parameter was

needed to be quantized. For a target composite data rate the proportioning of allocated bits to the various parameters, was optimised for best subjective quality.

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.

3. ADAPTIVE TRANSFORMATION

For a first order Markov process, the DCT has been reported to be asymptotically optimal as block size extends to infinity or adjacent correlation coefficient tends to unity [2][9]. For the segmented MELP parameters, however it was shown in [7] that the optimal KLT determined in a fixed average sense (fixed KLT) can improve the coding gain significantly over DCT. A fixed KLT scheme can be implemented by predetermining the transform in a training session and making it available at the encoder and decoder, thereby avoiding the KLT calculations and the transform encoding problems in the coding phase [7].

This fixed KLT diagonalizes the average covariance matrix determined over all data and for any given group of data blocks, however it does not diagonalise the local covariance matrix of that group of data blocks. For each group of data blocks, therefore a more optimal transform than this fixed KLT, exists and this locally optimal KLT can be determined via it's local covariance matrix.

One approach to avoid the expensive KLT calculations during the coding phase and to retain some of the benefit of locally optimal KLTs is to choose a transform from a codebook of possible transforms for the particular type of data blocks, using the vector quantization methodology. The transforms defined in the codebook would be chosen during a training session and referenced by a simple codebook index. The use of locally optimal KLTs however poses a new problem of quantizing the current transform information in addition to the transform output coefficients.

4. OPTIMIZATION CRITERION

To develop an optimization criterion for the transform quantizer, consider the scalar quantization of 2D, transform output coefficient block, $\theta_{MN}=\{\theta_{kl}\ ;\ k=1, 2..., M,\ l=1, 2..., N\}$. The resultant quantizer error variance for element θ_{kl} can be expressed as,

$$\sigma^2_{q\,kl} = c.\frac{\sigma^2_{\theta kl}}{2^{2R_{kl}}} \quad , \tag{1}$$

where $\sigma^2_{\theta kl}$ is the variance of θ_{kl}, R_{kl} is the number of bits used to quantize θ_{kl} and c is a constant known as the quantizer performance factor [2] and is same for all θ_{kl} if they each have the same PDF shape. Rearranging (1),

$$R_{kl} = \frac{1}{2}\log_2\left(\frac{\sigma^2_{\theta kl}}{\beta}\right) \quad \text{where} \quad \beta = \frac{\sigma^2_{qkl}}{c}. \tag{2}$$

For the same quantizer error variance for all elements of θ_{MN}, (Quantizer error variance should be maintained the same by a proper bit allocation so that for a given average bit rate the average quantizer noise is a minimum), the β value should be a constant. Therefore the average bit rate (per coefficient) for θ_{MN} can be written as,

$$R = \frac{1}{2}\log_2\left(\frac{GM}{\beta}\right), \text{ where } GM = \left(\prod_{k,l=1}^{M,N}\sigma^2_{\theta kl}\right)^{\frac{1}{MN}}. \tag{3}$$

Now consider the row wise (horizontal) 1D transformation of the training data block of M by N, $\mathbf{X}_{MN}=\{x_{mn}\ ;\ m=1, 2..., M,\ n=1, 2..., N\}$. If K different codebook transforms are used to map a total of L training data blocks into K different partitions of transform coefficient blocks, G^{θ}_r; r=1, 2..., K; categorised by the identity of the transform being taken, the total bit requirement to quantize all the transform coefficients of L training data blocks is,

$$R_{Total} = \sum_{r=1}^{K} M.N.L_r.(R)_{\theta r}\ , \tag{4}$$

where L_r is the number of data blocks coded by the r^{th} transform and $(R)_{\theta r}$ is the average bit rate, (3) for the r^{th} partition of transform coefficient blocks, G^{θ}_r.

Note that, $$L = \sum_{r=1}^{K} L_r\ .$$

Substituting (3) for $(R)_{\theta r}$ in (4), again assuming the same β value across all partitions of transform coefficient blocks, the total bit requirement can be written as,

$$R_{Total} = \frac{1}{2}\log_2\left(\frac{\prod_{r=1}^{K}\left[\prod_{k,l=1}^{M,N}\sigma^2_{\theta^r_{kl}}\right]^{L_r}}{\beta^{MNL}}\right), \tag{5}$$

where $\sigma^2_{\theta^r_{kl}}$ is the variance of the k^{th} transform coefficient of the r^{th} partition of transform coefficient blocks, G^{θ}_r. For the row wise transformation of $\mathbf{X}_{MN}$, assuming,

$$\sigma^2_{\theta^r_{kl}} = \sigma^2_{\theta^r_l} \quad \text{for} \quad 1 \le k \le M\ , \tag{6}$$

R_{Total}, (5) is a minimum when,

$$Z = \left[\prod_{l=1}^{N}\sigma^2_{\theta^1_l}\right]^{L_1}.\left[\prod_{l=1}^{N}\sigma^2_{\theta^2_l}\right]^{L_2}...\left[\prod_{l=1}^{N}\sigma^2_{\theta^K_l}\right]^{L_K}, \tag{7}$$

is a minimum.

From the optimal transform theory [2] it follows that,

$$\prod_{k=1}^{N} \sigma^2_{\theta^r_k} \geq \left|\mathbf{R}^r_{xx}\right|, \qquad (8)$$

where $\mathbf{R}^r_{xx}$ is the average covariance matrix of the data blocks corresponding to the r^{th} partition of transform coefficient blocks G^{θ}_r. If these data blocks are abbreviated by G^x_r, $\mathbf{R}^r_{xx}$ can also be written as,

$$\mathbf{R}^r_{xx} = \frac{1}{L_r} \sum_{\mathbf{X}_i \in G^x_r} \mathbf{R}_{xxi} \quad ; \quad \mathbf{R}_{xxi} = \frac{\mathbf{X}^T_{MNi} \cdot \mathbf{X}_{MNi}}{M}, \qquad (9)$$

where $\mathbf{R}_{xxi}$ is the local covariance matrix of the i^{th} training data block $\mathbf{X}_{MNi}$. These are Hermitian Positive Definite real matrices of dimensions N by N.

The lower bound of the product of transform coefficient variances can only be attained when the data blocks in G^x_r were transformed to G^{θ}_r by the KLT which was determined via $\mathbf{R}^r_{xx}$. This minimizes (7) for a given partitioning $\{G^x_r ; r=1, 2..., K\}$ and can be written as,

$$Z = \left|\mathbf{R}^1_{xx}\right|^{L_1} \cdot \left|\mathbf{R}^2_{xx}\right|^{L_2} \cdots \left|\mathbf{R}^K_{xx}\right|^{L_K} \qquad (10)$$

For a given K, ie. for the use of a fixed number of transforms, the bit requirement for encoding the index of the transform is fixed and hence the transform quantizer optimization problem reduces to partitioning of data blocks into G^x_r; r=1, 2..., K, to minimize (10).

5. SOLUTION METHODOLOGY

A direct analytical solution to optimize (10) is not available and the solution space is also extremely large to perform an exhaustive search. Therefore following an analogous argument to the GL algorithm for VQ design, the following solution methodology is proposed.

Step1. Begin with an initial arbitrary representative codebook of transforms.

Step2. Given a representative codebook of transforms find the optimal partitioning, $\{G^x_r ; r=1, 2..., K\}$.

Step3. Find the optimal representative transform codebook, for the partitioning just determined.

Step4. Evaluate performance for the new representative codebook and check for convergence to a final solution. if not, iterate the process from step 2.

The initial transform codebook can be generated by arbitrarily assigning the training data blocks to partitions and calculating the average KLTs for each partition.

The optimal 1D transform (row wise) for a single 2D data block of M by N is the KLT evaluated via it's local covariance matrix, $\mathbf{R}_{xxi}$ determined by (9), and it completely diagonalizes $\mathbf{R}_{xxi}$. When $\mathbf{R}_{xxi}$ is completely diagonalized the geometric mean of the transform coefficient variances (GM), evaluated locally as,

$$GM = \left[\prod_{l=1}^{N}\left(\frac{1}{M}\sum_{k=1}^{M}\theta^2_{kl}\right)\right]^{\frac{1}{N}}, \qquad (11)$$

is a minimum. Further more, for a sub-optimal transform, the smaller the value of GM the more diagonalized will be the $\mathbf{R}_{xxi}$. The selection criterion for optimal partitioning in step 2, can thus be chosen so as to achieve the minimum GM; GM as given by (11).

Following the argument of the previous section, for a given partitioning, the optimal representation for each partition, for step 3, is the average KLT calculated via the average covariance matrix, $\mathbf{R}^r_{xx}$ determined over the data blocks in that partition via (9).

Step 4 provides an exit condition for the iteration, when the algorithm has sufficiently converged to a solution.

6. EVALUATION

To evaluate the potential benefit of the proposed adaptive transformation scheme, the transform coding gain was calculated for the interframe and intraframe adaptive transformations of 2D LSP blocks resulted by the segmentation algorithm of section 2. Complete TIMIT training data base, lowpass filtered at 3.4 kHz and decimated to 8 kHz, was utilised for training of two transform codebooks each of size 8, for the interframe and intraframe transforms.

Transform coding gain is defined as,

$$G_{TC} = 10.\log_{10}\left[\frac{\sigma^2_{q,PCM}}{\sigma^2_{q,TC}}\right] \qquad (12)$$

where $\sigma^2_{q,PCM}$ and $\sigma^2_{q,TC}$ are the quantizer noise variances for the PCM and transform coding schems respectively.

Figures 1 and 2, show the transform coding gain in dB for different transformation schemes against the block size for the intraframe and interframe transformations respectively. Coding gains for the fixed transforms, DCT and fixed KLT are also plotted for the comparison. These graphs show only a 0.5 to 1dB improvement in coding gain for the adaptive KLT scheme over the fixed KLT scheme. This corresponds to a 1-2 bit savings per LSP set for the adaptive KLT scheme over the fixed KLT scheme for the same level of mean square error distortion.

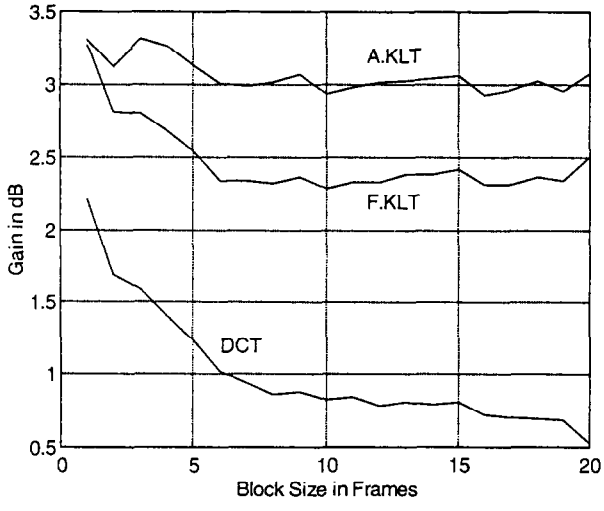

Figure 1. Intraframe transform coding gain for voiced LSP blocks.

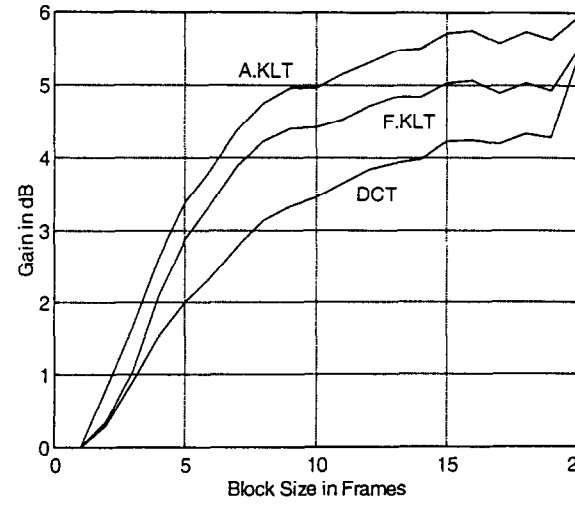

Figure 2. Interframe transform coding gain for voiced LSP blocks.

7. SUMMARY

A criterion for the optimization of a transform quantizer was developed considering the average bit rate requirement for the transform coefficient quantization. An iterative solution for the quantizer optimization was proposed. Preliminary evaluation of the adaptive transformation scheme in context of the coding scheme described in section 2, however indicated only a marginal improvement in the coding gain. The evaluation does not either account for the overhead of codebook index transmission.

The proposed adaptive transformation scheme can also be utilised, however in other applications where a dynamic transformation of data is beneficial.

Further research and testing is being presently carried out to investigate the optimality of the iterative solution of section 5 and for a complete evaluation of the proposed adaptive transformation scheme, both objectively and subjectively.

8. REFERENCES

[1] 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..

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

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

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

[5] 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.

[6] Mudugamuwa, D.J., Bradley, A.B., 1997, 'Adaptive transform coding for linear predictive residual', *5th European Conference on Speech Communication and Technology*, vol.1, pp433-436.

[7] Mudugamuwa, D.J., Bradley, A.B., 1998, 'Optimal transform for segmented parametric speech coding', *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, vol.I, pp53-56.

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

[9] Rao, K.R., Yip, P., 1990, *Discrete Cosine Transform-Algorithms Advantages and Applications*, Academic Press, Inc., Boston.