# OPTIMAL TRANSFORM 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, 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 in low data rate speech coding to exploit both the interframe and the intraframe correlation [9][2]. This paper investigates the potential for optimising the transform for segmented parametric representation of speech.

## 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 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.

In the first stage of this research, segmentation techniques together with the Discrete Cosine Transform (DCT) were employed for the MELP vocoder described in [6] and 50% bit rate reduction was achieved compared to the direct scalar quantized case, for the same level of subjective distortion [7].

In view of optimizing the above coding scheme, an optimal transformation of Line Spectral Frequencies (LSPs), within an effective segmentation frame work is investigated in this paper. Use of the optimal Karhunen-Loeve Transform (KLT), in a fixed average sense is investigated.

## 2. DCT 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 are 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 frames. The maximum block sizes were 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 transform compaction. Segmentation was implemented in a way to ensure no fragmentation of the blocks occur due to the limited buffer size.

For each vector parameter block a two dimensional discrete cosine transform (2D-DCT) 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.

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 [4][5] were designed using the probability density functions (pdf) obtained from the transformed coefficients themselves.

For each transform coefficient within a parameter, bit allocation was determined by it's variance, according to [3],[7] 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.

For the evaluation of the above coding scheme a direct scaler quantized version of the MELP coder, that does not use transform compaction to exploit the intraframe and interframe redundancies was also implemented. The final bit rates for the two coders; the DCT and the direct scaler quantized versions, were selected so that at a 95% confidence limit, listeners could not differentiate the quantized and unquantized versions of the synthesized speech output for 80% of the phrases selected from the TIMIT database. Results of these subjective tests have shown a need for 72 bits per frame for the direct scaler quantized case and 35 bits per frame for the DCT case to satisfy this criteria.

For the DCT coded MELP model, this represents an overall data rate of 1533 bits per second including segmental information overhead.

## 3. OPTIMAL TRANSFORM

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 [3],[8]. As the first step in optimizing the coding scheme described in section 2, the optimal transform for the LSPs was investigated. For LSP parameters the optimal transform across the frames (interframe) over a fixed non adaptive block size is very close to the DCT. However when the DCT was applied with the prescribed adaptive segmentation, it was found to perform well below the KLT, in either dimension. The poor performance of the DCT could be seen in two different perspectives: transform coding gain and the mismatch of frequency bands of the DCT with that of the KLT. Transform coding gain is defined as,

$$G_{TC} = 10.\log_{10}\left[\frac{AM}{GM}\right]$$

where AM and GM are arithmetic and geometric means of the variances of the transform coefficients.

The KLTs were estimated by solving the conventional eigen value problem [3],[8] of the interframe and intraframe covariance matrices for each different block size and type. Since the segmentation algorithm classifies the speech signal in to 48 (20 voiced, 8 unvoiced and 20 silence) different categories, for reliable estimation of correlation coefficients, the complete TIMIT training data base was utilised.

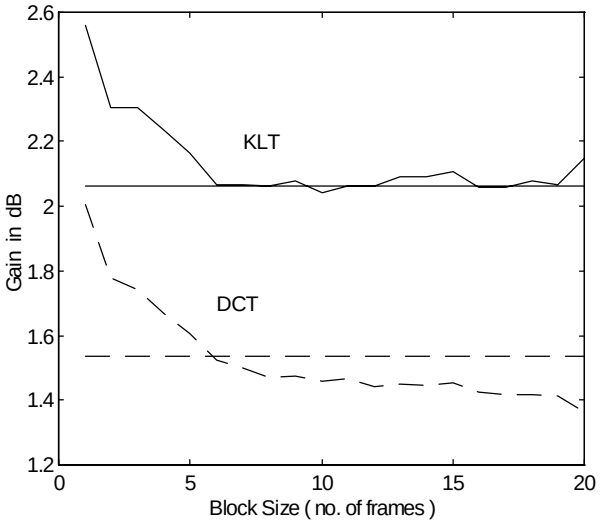


**Figure 1.** Intraframe transform coding gain for voiced blocks for the DCT and the KLT.

In figures 1 and 2, the transform coding gain in dB is plotted for both the DCT and the KLT against the different voiced block sizes for intraframe and interframe cases respectively. The two horizontal lines in figure 1 and the two curves marked as 'fixed' in figure 2 correspond to the transform coding gain that would be obtained with non-adaptive segmentation. ie. all speech is segmented for a fixed block size and voiced and unvoiced parts are allowed to be mixed.

### 3.1 Intraframe Transform

From figure 1 it is clear that for the intraframe transform, the DCT performs about 2.5 dB below the optimal (KLT) even with fixed segmentation (two horizontal lines). This corresponds to a 2-3 bits saving per frame for the KLT over the DCT for the same distortion level. Figure 3 shows the distribution of the first five frequency bands of the intraframe KLT estimated with fixed segmentation. These frequency bands correspond to the basis functions (eigen vectors) associated with highest eigen values. The remaining five bands are not shown for clarity and their shapes are also found to have little significance for the entropy reduction. Clearly the uniform band structure of the DCT is significantly different from that of figure 3.

The adaptive segmentation had little effect on intraframe transform coding gain though for the KLT a small nett

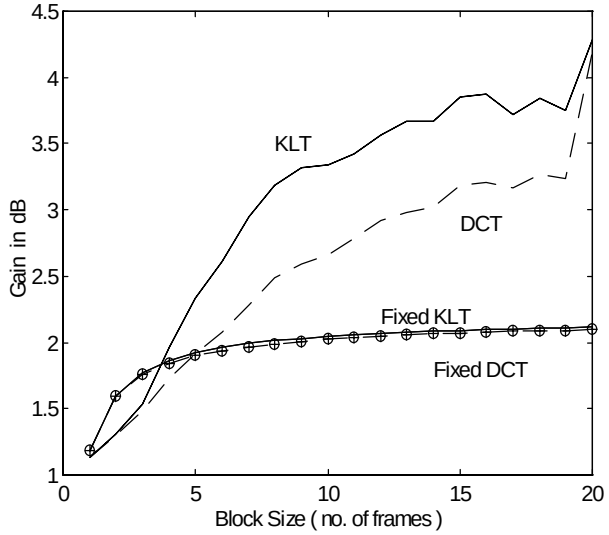


**Figure 2.** Interframe transform coding gain for voiced blocks for the DCT and the KLT.

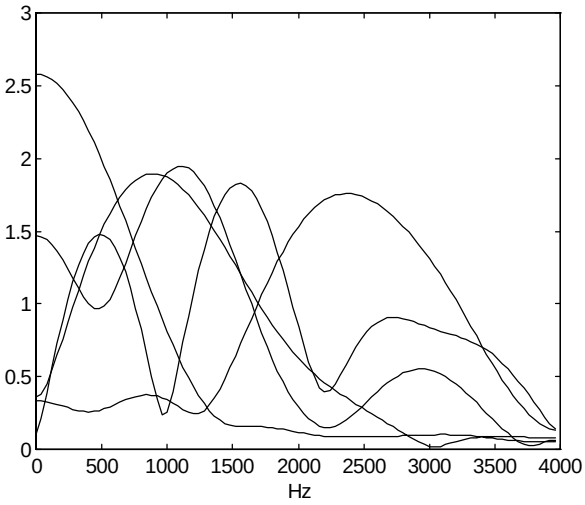


**Figure 3.** Principal five frequency bands of intraframe KLT with fixed segmentation.

improvement in gain could be observed. It should also be noted that the overall proportion of frames in small blocks is much fewer than larger blocks. For the unvoiced speech case, the gain was observed to be above the two horizontal lines of figure 1 for both the DCT and the KLT. The trend in the gain for DCT and KLT for the segmented LSPs suggests, that LSP coefficients in small blocks are more correlated within a frame compared to larger blocks.

It could also be observed though not clearly apparent in figure 1, that the difference between the KLT and DCT is becoming narrower when the segmentation is poor. Due to the limitation on maximum block size, blocks larger than

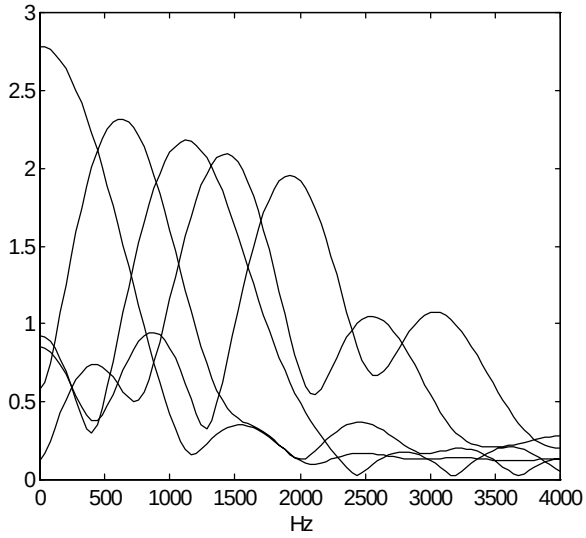


**Figure 4.** Principal five frequency bands of intraframe KLT for single frame voiced blocks

the maximum size, get fragmented and fall back mostly to the smaller blocks. For single frame voiced blocks this is as high as 30% of the blocks. As a result, for smaller blocks the KLT band structure is becoming closer to the uniform band structure of the DCT. This is clearly illustrated in figure 4, where the principle five frequency bands of intraframe KLT for single frame voiced blocks are plotted.

## 3.2 Interframe Transform

With fixed segmentation the DCT across the frames (interframe) performs almost optimally as shown in figure 2, 'mixed' case. But clearly adaptive segmentation boosts the transform coding gain for both DCT and KLT. In this case the performance of the KLT over the DCT indicates bit savings of up to about 1 bit per LSP set for the same level of distortion. These differences of the DCT from the KLT could also be observed in the frequency responses of the basis functions.

It can be seen from figure 2 that for up to 4 frame blocks, coding gain for the adaptive segmented case is less than that with fixed segmentation. This was also true for unvoiced blocks. This suggests that the correlation of LSPs across the adjacent frames is less for smaller blocks than that for larger blocks.

Experiments were carried out to determine the benefit of using individually optimised transforms for each row and column for each block size and type of LSPs. At the expense of considerable complexity an additional benefit of only 0.2 - 0.4 bits per frame could be obtained for the same level of distortion.

## 4. QUANTIZATION RESULTS

The LSP coding scheme in section 3, with the adaptive segmentation and with fixed optimal transforms for each block size and type, was implemented. The normalised KLT coefficients were quantized using pdf optimised scalar quantizers [4][5] as in section 2. To evaluate the overall performance of this scheme, average spectral distortion was evaluated across all test frames. The spectral distortion (SD) on a frame basis is defined as,

$$SD = \left[ \frac{100}{2\pi} \int_{-\pi}^{\pi} \left| \log_{10}(A(\omega)) - \log_{10}(A'(\omega)) \right|^2 d\omega \right]^{\frac{1}{2}}$$

where A(ω) and A'(ω) are the unquantized and quantized spectral responses of the LP filter. The TIMIT test speech database, low pass filtered at 3.4 KHz and decimated to 8 KHz, was used.

**Table 1.** Objective results of the DCT quantization

| Bits/frm | Avg. SD | % >2 dB | % >4 dB |
|---|---|---|---|
| 14.25 | 1.152 | 5.0 | 0.3 |
| 15.59 | 1.093 | 4.1 | 0.5 |
| 16.75 | 1.034 | 3.6 | 0.4 |
| 18.00 | 0.975 | 3.4 | 0.4 |

**Table 2.** Objective results of the KLT quantization

| Bits/frm | Avg. SD | % >2 dB | % >4 dB |
|---|---|---|---|
| 11.81 | 1.129 | 3.4 | 0.1 |
| 13.05 | 1.068 | 2.8 | 0.2 |
| 14.20 | 1.002 | 2.3 | 0.2 |
| 15.29 | 0.950 | 1.9 | 0.1 |

Tables 1 and 2 provide the average SD, calculated across 30,000 voiced and unvoiced frames for the DCT and the KLT quantization schemes respectively. The DCT coding scheme of section 2, was applied to contiguous speech with voiced, unvoiced and silence blocks included. The results of tables 1 and 2 do not benefit from the inclusion of silence blocks.

## 5. SUMMARY

In view of improving the coding scheme in section 2, optimal transformation for the segmented LSPs was studied. Estimations of the bit savings via the variances of transform coefficients, in section 3, were justified in section 4. Use of the fixed optimal transforms with the adaptive segmentation permits the quantization of LP filter using 14 bits per frame (22.5 msec) at 1 dB spectral distortion, at the expense of 450 msec coding delay. This is a benefit of about 3 bits per frame over DCT at no additional overhead or computational cost. The quantization of the transform coefficients can further be improved using vector quantization techniques.

Further research is presently being carried out towards the development of an adaptive transformation scheme with locally optimal transforms in place of the non-adaptive globally optimal KLTs.

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

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

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

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

[6] 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.

[7] 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.

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

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