Proceedings of The Institute of Acoustics

ADAPTIVE TRANSFORM CODING OF SPEECH AT 9.6 kB/S AND BELOW

J. M. Rye, and S. C. Dupree

Joint Speech Research Unit, Cheltenham, Glos, GL52 5AJ.

INTRODUCTION

Over the last few years considerable interest has been shown in adaptive transform coding (ATC) as a method of coding speech signals at medium bit-rates, (1,2). Fig. 1 shows a general adaptive transform coder. The input speech signal is split into time segments, or frames, typically of 16-32 ms duration. The mean power of the signal in each frame is measured and used to normalise each frame to unity variance. The frame of data is then transformed to the frequency domain in an attempt to orthogonalise the data. The transform coefficients are quantised and transmitted to the receiver, the number of bits used for each coefficient being varied from frame to frame. The allocation of bits is controlled by a simplified description of each frame which, along with the overall gain, is transmitted as "side information". The receiver, Fig. 1b, uses the side information to interpret the coefficient data bits. The de-coded coefficients are then inverse transformed and de-normalised to reconstruct the time signal.

Various ATC algorithms have been suggested and implemented, differing mainly in the choice of side information system. Our work was begun as an investigation into ATC, from which it was hoped to be able to make good choices for the side information system, based on quality for a given data rate, but trying to restrict complexity so that a hardware implementation would be possible in the near future. All simulations were made using non real-time Fortran programs running on a PDP 11/40 minicomputer, using an FPS AP-120B array processor for the computationally expensive operations.

PROBLEMS WITH ATC AT LOW RATES

At rates above about 16 kb/s a simple ATC scheme as described in (1) gives good quality output. As the bit-rate is reduced so the coefficient quantisers and the side information coder become starved of bits, leading to the following observable effects.

a) The shortage of bits in the coefficient quantisers leads to zero bits being allocated to lower energy spectral regions. During voiced speech these regions tend to be located at higher frequencies, causing the coded signal to become deficient in high frequency energy. During unvoiced sounds the effect is of course reversed. However the voiced portions dominate, giving a muffled quality to the speech output. We choose to call this particular distortion "zero-bit effect".

b) The output speech contains rapidly varying tone-like distortions or "burbling" noises, thought to be caused by intermittent allocation of bits to isolated coefficients or groups of coefficients. At low bit-rates the effect is more noticeable because the coded speech spectrum becomes quite broken up and the ear is able to perceive the intermittent tones as separate from the coded speech signal (especially tones isolated by a critical band or more).

c) Frame boundary miss-match effects are worse due to there being more error in each frame.

d) The reduction of bits available to the side information scheme means that the bits for the coefficient quantisers are not well allocated, worsening to varying degrees the effects a), b) and c).

DETAILS OF ATC IMPLEMENTATION

To minimise some of the problems of low bit-rate ATC, we tried to take account of important aspects of speech production and perception.

The Transform

The transform we use is the discrete cosine transform (DCT) which can be defined as the real part of the DFT of an even extension (folding) of the time signal. Various reasons have been given for using a DCT ranging from purely theoretical arguments (1) to more practical considerations such as boundary effects (2). However the DCT does partially obscure the harmonic properties of the voiced speech spectrum due to the inherent folding.

The Side-Information

Our side information represents both the envelope structure and any important fine structure present in the DCT coefficients. We felt it important to have direct control over the allocation of bits, and thus the distortion, in each critical band (3). The envelope side information therefore consists of a code representing the power in each of 19 approximate critical bands between 0 and 4 kHz. The envelope shape is coded using a two bit difference coder similar to that in (4), except that five bits are used to code the overall height of the spectrum. The envelope information, including the overall gain uses 46 bits per frame.

The short time spectrum of voiced speech sounds is known to exhibit harmonic structure. For these sounds it would be reasonable to allocate bits primarily to the speech harmonics. Obviously, a more sophisticated side information scheme is then needed to send this harmonic detail. We chose to design a rather general fine structure extractor coding selected coefficients from the DCT cepstrum (the cosine transform of the log of the transform coefficients). A DCT cepstrum is convenient as it uses building blocks already available in the ATC coder. The low order cepstral coefficients are ignored because they contain information already described by the envelope side information. The largest positive peak in the lower half of the cepstrum is then coded with two bits for its value and seven for its index (for a frame of 256 samples). The largest of the three coefficients centred on double the index of this "fundamental" peak is coded with three bits. When the side information is interpreted its index is assumed to be precisely double the "fundamental". Of the remaining coefficients, the largest two are each coded with eight bits for index, and three bits for value. The levels for the quantiser were chosen by applying the theory of Max (5) to the amplitude distribution of the selected cepstral coefficients obtained from a few seconds of male and female speech. The fine structure side information requires a total of 34 bits per frame.

Bit Allocation in Frequency

The number of bits allocated to each coefficient is calculated from the basis of rate distortion theory in a similar fashion to that in (1,2,6). If $s(i),(i=1,L)$ are the side information estimates of the log-base-2 DCT coefficients, and we allow the resulting noise log-spectrum to be a proportion, r, of s(i), then the theory gives a bit allocation:

$$b(i) = B/L + c(1-r)[s(i) - (1/L)\sum_{1}^{L} s(i)] \qquad (1$$

where B is the total number of bits per frame and L is the frame length in samples. The term, c, often assumed to be unity, is in fact a function of s(i) as pointed out by Berouti and Makhoul (7). We found a minimum distortion level corresponded to an average value for c of about 1.3. We achieve different noise shaping for the envelope and fine structure by separately scaling each by (1-r), using values for r of 0.1 and -0.25 respectively. This effectively brings up the noise under the envelope peaks and suppresses the noise on harmonic peaks. The two scaled spectra are combined by addition, with an offset calculated for each critical band, to ensure that the new estimate of the total power in the band is equal to that of the envelope estimate alone. We also use the noise masking depths given in (8) to derive an upper limit to the bit allocation per coefficient, which is different in each critical band.

Bit Allocation in Time

Hearing research has indicated that the masking of a quiet signal by an adjacent louder signal lasts for a few tens of milliseconds for signals (tones and noise) in the same critical band (9). Thus a louder speech signal could contribute to the masking of quantising noise in an adjacent low intensity region. We employ a buffer of about 3 frames length and distribute the total bits between the frames depending on the power in each frame in a similar fashion to that in (6).

Quantising the Transform Coefficients

Each DCT coefficient, y(i), is gain normalised and then quantised with a unit variance quantiser. The normalising gain, $1/\hat{y}(i)$, is the reciprocal of the side information estimate of the DCT amplitude spectrum. We calculate the non-linear quantiser levels by applying the theory of Max to the long term amplitude distribution of $y(i)/\hat{y}(i)$, obtained from male and female speech. It is important here to note that this distribution is dependent on the quality of the side information. For a simple scheme as in (1), it is approximately gaussian. If the side information gave a perfect estimate of the DCT coefficient amplitudes, then the distribution would be a pair of delta functions at +/- 1. The distribution we found is shown in Fig 2. Using levels calculated from this distribution rather than those from a gaussian slightly improved the coded speech quality.

The Receiver

In the receiver the side-information is used to interpret the coefficient bit-stream. The coefficients are then transformed by the inverse DCT (IDCT), and the resultant time signal multiplied by the overall gain. We apply a trapezoidal window to each frame (such that the sum of two overlapping windows is always unity) and then overlap-add before output.

CONCLUSIONS

In our simulations we have attempted to code the maximum available bandwidth of speech (0-4 kHz bandwidth for 8 kHz sampling), and have found that ATC can give good quality output at rates of the order of 9.6 kb/s. We believe the following points to be important in producing good quality ATC at low rates.

a) A critical band structure (3,8) should be used in the side information scheme, allowing independent control of distortion in regions that are perceptually separable.

b) The side information should also be able to model any important fine structure in the transform coefficients. Hard decisions of periodicity and voicing should be avoided.

c) The maximum number of bits available to each coefficient should be limited to reflect approximately the variation with frequency of the noise masking abilities of the ear (8).

d) Noise shaping is useful in that it allows a trade-off between zero-bit effects and burbling noise on the one hand, and overall roughness on the other.

e) The adaptive allocation of bits among successive frames may be useful at very low-rates where there would not otherwise be enough bits to code the louder frames sufficiently well.

f) The levels for all quantisers should be matched to the observed distribution of the signal to be quantised, using the theory in (5).

REFERENCES

1. R. ZELINSKI and P. NOLL 1977 IEEE Trans. Acoust., Speech and Signal Processing, ASSP-25, 299-309. Adaptive Transform Coding of Speech Signals.

2. J. M. TRIBOLET and R. E. CROCHIERE 1979 IEEE Trans. Acoust., Speech and Signal Processing ASSP-27, 512-530. Frequency Domain Coding of Speech.

3. B. SCHARF 1970 in Foundations of Modern Auditory Theory ed. J. V. Tobias 1,159-202. Academic Press, London and New York. Critical Bands.

4. J. N. HOLMES 1980 IEE Proc. 127 pt. F, 53-60. The J.S.R.U. Channel Vocoder.

5. J. MAX 1960 IRE Trans. IT-6, 7-12. Quantising for Minimum Distortion.

6. J. MAKHOUL, M. BEROUTI, and M. A. KRASNER 1981 Proc. IEEE ICASSP, 611-614. Time and Frequency Domain Noise Shaping in Speech Coding.

7. M. BEROUTI and J. MAKHOUL 1980 IEEE ICASSP, 356-359. An Embedded-Code, Multirate Speech Transform Coder.

8. M. A. KRASNER 1979 Speech Communication Papers presented at the 97th meeting of A.S.A., 381-384. Digital Encoding of Speech Based on the Properties of the Human Auditory System.

9. H. FASTL 1977 Acustica 36, 326-328. Temporal Masking Effects II. Critical Band Noise Masker.

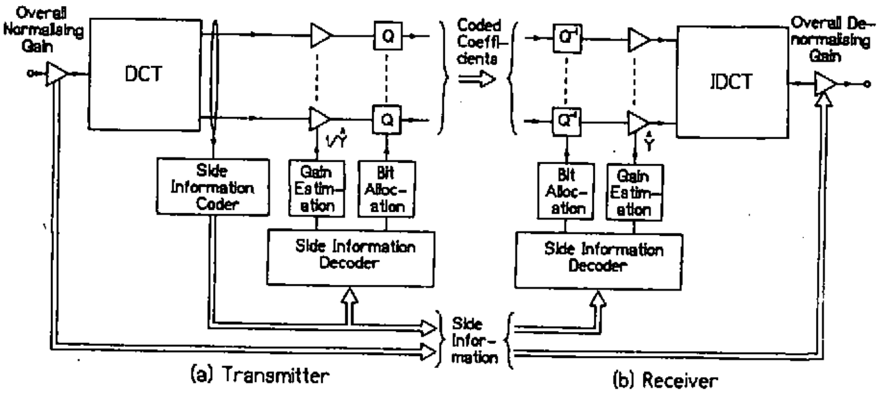

Fig. 1 An Adaptive Transform Coder

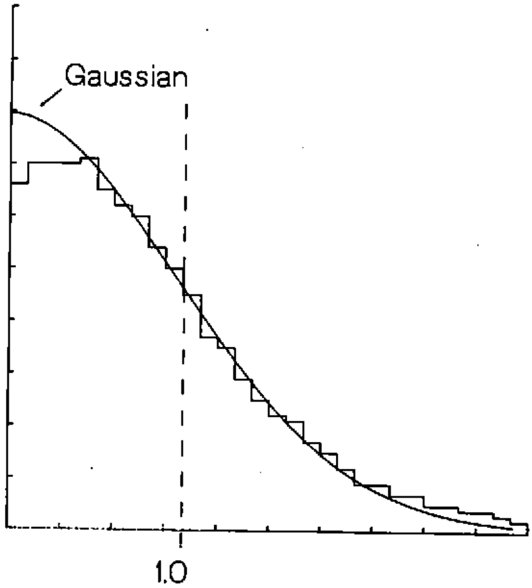

Fig. 2 Comparison of a Gaussian Distribution with the Amplitude Distribution of Gain Normalised DCT Coeffs., $y(i)/\hat{y}(i)$.