Perceptual quantisation of LPC excitation parameters

W.-W.Chang
D.-Y.Wang

Indexing terms: Perceptual quantisation, Digital audio compression, LPC coder

Abstract: A new approach to wideband digital audio compression at 64 kbit/s is presented. This LPC coder features a sinusoidal excitation model and adaptive bit-allocation based on a mask-to-noise ratio perceptual measure. Owing to its localised spectral sensitivity, sinusoidal excitation representation can be shown to provide an ideal framework for incorporating an estimated masking threshold in the design of noise spectral shaping. Quantisation of LPC and excitation parameters is also examined. Simulation results indicate that sine waves are preferred to spectrally flat signals for use in excitation modelling, because in the former the estimated masking threshold can be more precisely implemented in the distribution of quantisation noise.

1 Introduction

With recent developments in coding technology, high quality digital audio transmission over 64 kbit/s ISDN channels has become feasible [1]. In essence, perceptual coding systems are designed to use statistical correlation to remove redundancies, and also to eliminate the perceptual irrelevancy by applying psychoacoustic measures. Most studies have concentrated on transform coding [2] and subband coding [3], in which audio spectra are subdivided into critical bands and then quantised in accordance with the estimated masking threshold. An alternative approach to audio representation is based on the linear predictive coding (LPC) model [4, 5], which considers waveforms to be outputs from an all-pole filter that uses spectrally flat (Gaussian white noise for unvoiced signals and multiple impulses for voiced signals) excitation signals. Though LPC models have been widely used in audio coding, a few comments can still be made concerning the further enhancements. First, analysis of experimental data shows that real residual signals exhibit predominantly pulselike trends in the frequency domain, which contrasts sharply with signals from spectrally flat excitation sources, such as multipulse LPC (MPLPC) coders. Secondly, most psychoacoustic experiment results are expressed in the frequency domain and hence are not directly applicable to LPC models. Thus, many approaches have been considered which attempt to benefit from perceptual masking of quantising noise in LPC-based coders [6]. Early attempts employed relatively simple techniques for incorporating the masking properties either in postfiltering [7] or in noise feedback coding [8]. Recently proposed analysis-by-synthesis LPC coders utilise the prediction coefficients [9] or the masking threshold [10] to implement perceptual weighting filters for excitation searches. Further improvement can only be realised through some intelligent exploitation of new findings in excitation representation.

The strategy applied here is to represent excitation waveforms as a sum of sine waves with arbitrary frequencies, amplitudes and phases. This allows us to consider each sine wave as a separate spectral line, and proves to be advantageous in perceptual coding applications. From the perspective of noise-shaping, this sinusoidal representation provides an ideal framework for incorporating perceptual informations, since individual sinusoids can be independently quantised without the leakage of quantisation noise from one spectral line to another. This error localisation property also helps in developing the dynamic bit allocation required for quantisation of excitation parameters. The concept of sinusoidal representation was originally developed to provide an approximation of speech waveforms [11]. However, this straightforward approach tends to cause parameter discontinuities at frame boundaries leading to audible artifacts in steady-to-transient regions. As shortcomings become apparent, frequency tracking and parameter smoothing techniques must be introduced to deal with rapid changes during transient periods. As we shall see, the parameter continuity problem would not be a serious obstacle if sinusoidal analysis were performed on residual signals instead of on the incoming sound, as in [11].

2 Multisinusoid LPC coders

Most analysis-by-synthesis LPC coders decompose signals into the product of excitation and system spectra, and then represent the excitation by means of spectrally flat signals. However, from inspection of Fig. 1 it is evident that real residual signals tend to exhibit pulselike trends in the frequency domain. To provide better matches with peaky residual spectra, we propose to represent excitation waveforms as sums of sine waves with arbitrary amplitudes, frequencies and phases. Viewed from this perspective, the general form of a

IEE Proceedings online no. 19971676

Paper first received 11th October 1996 and in revised form 22nd September 1997

The authors are with the Department of Communication Engineering, National Chiao Tung University, Hsinchu, Taiwan, Republic of China

multisinusoid excitation model is given by

$$e(n) = \sum_{i=1}^{M} e_i(n) = \sum_{i=1}^{M} r_i \cos(\omega_i nT + \phi_i) \quad 1 \le n \le N \quad (1)$$

where N is the subframe length, M is the number of sinusoids, and the r_i, ω_i and ϕ_i represent amplitude, frequency, and phase respectively, of the ith sinusoidal component $e_i(n)$. Fig. 2 illustrates the functional block diagram of the proposed Multisinusoid LPC (MSLPC) encoder.

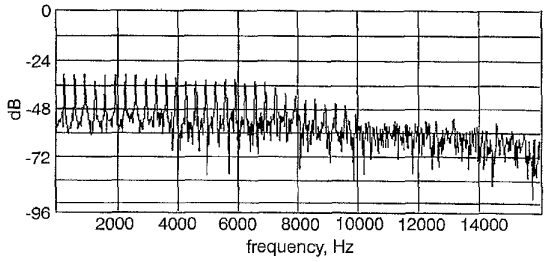

Fig. 1 *Spectrum of the original residual segment*

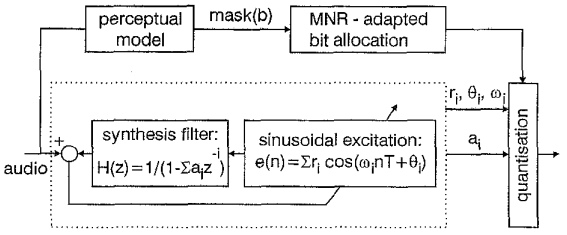

Fig. 2 *Block diagram of the multisinusoid-excited LPC encoder*
Synthesis filter: $H(z) = 1/(1 - \Sigma a_i z^{-i})$
Sinusoidal excitation: $e(n) = \Sigma r_i \cos(\omega_i nT + \theta_i)$

The proposed system performs psychoacoustic control of quantisation noise by using adaptive quantisation instead of the noise-weighting filters for excitation search, as do conventional analysis-by-synthesis LPC-based coders. Two basic types of system parameter can be identified: LPC parameters and excitation parameters. The LPC analysis is performed with the autocorrelation method once per frame, whereas excitation parameters are updated once per subframe. In our study, audio signals with a bandwidth of 15kHz were sampled at 32kHz and then segmented into frames of 300 samples long. Each frame was further divided into 6 subframes. Letting $h(n)$ denote the impulse response of the synthesis filter, we produce output signals $y(n)$ by taking the convolutional sum

$$y(n) = \sum_{i=1}^{M} [\alpha_i h_{ci}(n) + \beta_i h_{si}(n)] \quad 1 \le n \le N \quad (2)$$

where $\alpha_i = r_i \cos\phi_i$, $\beta_i = -r_i \sin\phi_i$, $h_{ci}(n) = \cos(\omega_i nT) * h(n)$, and $h_{si}(n) = \sin(\omega_i nT) * h(n)$.

Accurate identification of excitation parameters was considered to be the basis for the success of MSLPC. This was accomplished by minimising the squared-error distortion between the original signal $x(n)$ and the output signal $y(n)$. This minimisation process resulted in the matrix form of

$$\mathbf{S} \cdot \mathbf{g} = \mathbf{c} \quad (3)$$

where the entries in $\mathbf{g}$, $\mathbf{c}$ and $\mathbf{S}$ are given as follows for $1 \le j \le 2M$ and $1 \le k \le 2M$ respectively:

$$g_j = \begin{cases} \alpha_{(j+1)/2} & j \text{ odd} \\ \beta_{j/2} & j \text{ even} \end{cases} \quad (4)$$

$$c_j = \begin{cases} \mathbf{x} \cdot \mathbf{h}^T_{c(j+1)/2} & j \text{ odd} \\ \mathbf{x} \cdot \mathbf{h}^T_{s(j/2)} & j \text{ even} \end{cases} \quad (5)$$

$$S_{jk} = \begin{cases} \mathbf{h}_{c(j+1)/2} \cdot \mathbf{h}^T_{c(k+1)/2} & j,k \text{ odd} \\ \mathbf{h}_{s(j/2)} \cdot \mathbf{h}^T_{s(k/2)} & j,k \text{ even} \\ \mathbf{h}_{c(j+1)/2} \cdot \mathbf{h}^T_{s(k/2)} & j \text{ odd}, k \text{ even} \\ \mathbf{h}_{s(j/2)} \cdot \mathbf{h}^T_{c(k+1)/2} & j \text{ even}, k \text{ odd} \end{cases} \quad (6)$$

Using the Cholesky factorisation theorem [12], eqn. 3 can be solved more efficiently by decomposing the symmetric matrix $\mathbf{S}$ into the form of $\mathbf{GG}^{\mathrm{T}}$ where $\mathbf{G}$ is a lower triangular matrix with non-zero entries as follows:

$$G_{jj} = \sqrt{S_{jj} - \sum_{k=1}^{j-1} G_{jk}^2} \quad 1 \le j \le 2M \quad (7)$$

$$G_{jk} = \left(S_{jk} - \sum_{l=1}^{k-1} G_{jl} G_{kl}\right) \Big/ G_{kk} \quad 1 \le k \le j-1 \quad (8)$$

Proceeding in this way, we can rewrite eqn. 3 as follows:

$$\mathbf{Gq} = \mathbf{c} \quad (9)$$

$$\mathbf{G}^T\mathbf{g} = \mathbf{q} \quad (10)$$

where the entries in $\mathbf{q}$ are given by

$$q_j = \left(c_j - \sum_{k=1}^{j-1} G_{jk} q_k\right) \Big/ G_{jj} \quad 1 \le j \le 2M \quad (11)$$

Using this notation, the least squared error distortion is given by

$$E_{min}^{(M)} = E_{min}^{(M-1)} - \left(q_{2M-1}^2 + q_{2M}^2\right) \quad (12)$$

From inspection of eqn. 12, it is evident that the optimum values of the parameters $\{\omega_i\}$ and $\{r_i, \phi_i\}$ can be independently estimated. As regards the frequencies, a set of L candidates was chosen once per frame by locating the predominant peaks inherent in the associated audio spectrum. Next, only these L candidates were examined to find the M best frequencies needed within each of its 6 constituent subframes. Towards this end, the frequency of the ith component sine wave was taken as the location of the particular candidate, maximising the term $(q_{2i-1}^2 + q_{2i}^2)$. Once the frequencies were determined, the optimal values of $\{r_i, \phi_i\}$, which are exclusively embedded in $\mathbf{g}$, could be found by solving eqn. 10.

3 Quantisation and bit allocation

The next step in the present investigation is concerned with efficient quantisation of LPC and excitation parameters. There are numerous possible quantisation methods for these parameters to be transmitted at the rate of 64 kbit/s. Using an analysis frame of 300 sam-

ples, the total number of bits available per frame is 600. We have performed limited experiments with the perceptual quantisation approaches and found that the bit allocations in Table 1 provide good results in the case of $L = 13$ and $M = 7$, although they are not optimal. The quantisation aspects of the various parameters are detailed below.

Table 1: Bit allocation for MSLPC coders at 64kbit/s

Amplitudes	35 × 6
Phases	35 × 6
Frequencies	11 × 6
Frequency candidates	78
Maximum amplitude	5
LPC parameters	24
Bit allocation information	7
Total bits per frame	600

3.1 LPC parameters

For our study, a 10th order LPC analysis was chosen to characterise the spectral envelope information of the incoming sound. Prior to transmission, these LPC parameters were transformed into line spectral frequencies (LSFs) and then quantised using split vector quantisation at 24 bits/frame [13]. More explicitly, we divided the vectors of 10 LSFs into two parts: one consisting of the first 4 LSFs and the other consisting of the remaining 6 LSFs. Each of these two parts was allocated 12 bits. We first examined whether LSF parameters could be efficiently quantised using split-vector quantisation. The monophonic audio database for these studies consisted of 200 seconds of audio signals recorded from various musical instruments. The first 170 seconds of music was used for training, and the last 30 seconds of music was used for testing. The performance was evaluated in terms of spectral distortion (SD), defined as the root mean square difference between the original LPC log-power spectrum and its quantised version. An average SD of 1dB is usually accepted as the difference margin for spectral transparency. Since no SD score exceeded 1dB for any of our test samples, we can conclude that split-vector quantisation quantises LPC parameters at 24 bits/frame with transparent quality.

3.2 Excitation parameters

In the proposed MSLPC system, excitation parameters consist of the frequencies, amplitudes and phases of the component sine waves. As mentioned earlier, a set of 13 spectral peaks per frame were first located as candidates and then examined to find the 7 best frequencies needed within each of its 6 constituent subframes. In the range of 1024 DFT coefficients, a direct approach to representing each candidate position required the use of 9 bits. However, bit reduction can be aided by taking advantage of the statistical distribution of predominat peaks inherent in the audio spectrum. More specifically, we encoded the first candidate as an absolute location in the frame and the remaining candidates as differences from the previous one. To advance with this, 13 candidates were differentially encoded in 78 bits according to the bit allocation (5, 6, 6, 6, 6, 6, 6 ,6, 6, 6, 6, 6, 7). Since the 15 kHz bandwidth is resolved in a 1024-point DFT, this arrangement refelcts the observation that the first spectral peak is below 1kHz and the differences between successive peaks are usually within 2kHz, except that the last difference might exceed 4kHz. Next, we employed an enumerative source coding technique [14] to encode the 7 frequencies once per subframe. Note that the number of different possibilities involved in choosing 7 out of 13 candidates is given by C_7^{13}. Thus, the minimum number of bits required to code all possible patterns within a subframe is 11.

To quantise the amplitudes, an adaptive quantiser whose levels were adjusted to the maximum absolute value within a frame was used. This maximum absolute value, denoted by r_{max}, was logarithmically encoded in 5 bits. The individual amplitudes were then scaled and uniformly quantised using different degrees of bit resolution. The aim was to obtain a larger margin between the coder generated noise level and the audibility threshold of such artifacts. Following the work described in [2], we implemented a perceptual model to obtain the input parameters (mask-to-noise ratios) required to optimise the bit-rate adjustment procedure. The calculation started with a precise spectral analysis on 1024 windowed audio samples to generate its magnitude spectrum. The spectral lines were then examined to discriminate between tonelike and noiselike maskers by taking the spectral flatness measure as an indicator of tonality. Using rules known from psychoacoustics, the spread Bark spectrum was then calculated dependent on frequency position, loudness level and the nature of tonality. Finally, we obtained a vector of 24 masking thresholds, denoted by $\{mask(b), b = 1, 2, ..., 24\}$, from the spread Bark spectrum and from the absolute threshold in quiet. Each of them represents the maximum level of unnoticable quantisation noise in one critical band.

Constrained to producing a constant bit-rate for each frame, we proposed a dynamic bit allocation routine based on the MNR (mask-to-noise ratio) perceptual measure, which is defined as the ratio of the estimated masking threshold to actual coding noise. The primary goal was to minimise the total mask-to-noise ratio over each subframe by increasing the quantiser resolution for perceptually more important sinusoids until the number of bits available was exhausted. Let us assume that σ_i^2 is the variance of the ith filtered sinusoidal component, denoted by $s_i(n) = e_i(n) * h(n)$, and let σ_{qi}^2 denote the quantisation noise variance associated with R_i-bit uniform quantisation of the signal $s_i(n)$. The proposed bit allocation routine is an iterative procedure, where in each iteration the following steps proceed until all 35 bits have been allocated in coding the amplitudes:

(i) Calculate the error variances of all the sinusoids $1 \leq i \leq 7$

$$\sigma_{qi}^2 = \epsilon\sigma_i^2 / 2^{2R_i} \tag{13}$$

where ε is the corresponding quantiser performance factor

(ii) Calculate the MNR of all the sinusoids $1 \leq i \leq 7$

$$MNR(i) = mask(b) - 10\log_{10}\sigma_{q_i}^2 \quad \omega l_b \leq \omega_i \leq \omega h_b \tag{14}$$

where ωl_b and ωh_b denote, respectively, the lower and the upper boundaries of the bth critical band

(iii) Assign one additional bit to the particular sinusoid with the minimum MNR.

Once the final bit allocation was determined, the individual amplitudes were then uniformly quantised in the range of $[0, r_{max}]$ with different quantiser resolutions. As regards the phases, each sine wave was equally allocated 5 bits and uniformly quantised in the range 0 to 2π.

3.3 Side information description

To account for variabilities in input signal variances, we need to transmit some type of side information regarding the adaptation of bit allocation to changing MNR characteristics. Using this side information, the receiver can recover the bit allocation in the same way as the transmitter, thus making lengthy bit-rate calculations unnecessary. The side information in this coder includes a series of 7-bit pointers, each pointer corresponding one subframe. The decoding of bit allocation is done in a two-step procedure, first step consisting of reading 7 bits of information from the bit stream, which are then interpreted as an unsigned integer. The second step uses this integer as the index to a relevant table from which a 7-digit pattern, whose ith digit gives the number of bits allocated to quantize the ith sinusoidal amplitude is obtained. In this experiment, the table consisted of 128 entries empirically determined from the most frequently used bit allocations.

4 Experimental results

Computer simulations were conducted to examine the suitability of MNR-adapted bit allocation for use with the proposed multisinusoid LPC audio coder. The monophonic audio database for these studies consisted of electrified instrumental music, an oboe plus a piano, an orchestra, and sentential utterances spoken by two females, each 10 seconds in duration and sampled at 32kHz. First of all, a preliminary experiment was performed to examine the dependence of the performance gain on the number of sinusoids used to approximate the excitation waveform. Our general conclusion is that further improvement can be obtained by increasing the number of sinusoids. This is clearly shown in Fig. 3, in which the performance is evaluated in terms of segmental SNR (SNRSEG). However, the increasing demand for sinusoids tends to render parametric extraction so unwieldy as to make implementation impossible. Recognising this, we empirically chose $L = 13$ and $M = 7$ as the best compromise between coding gain and implementational complexity.

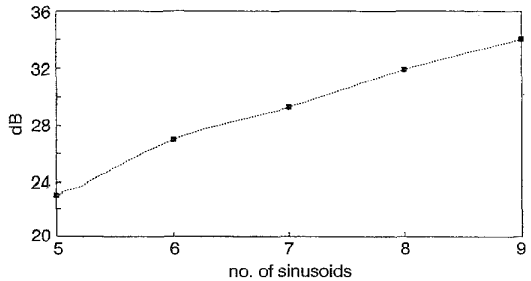

Fig. 3 *SNRSEG as a function of sinusoid density for multisinusoid coders*

The next problem addressed concerned the performance gain attainable by incorporating a multisinusoidal excitation model in the design of analysis-by-synthesis predictive coders. Table 2 shows the comparative performance results for 64 kbit/s coding of audio in conjunction with multipulse and multisinusoid based excitation. The reference MPLPC algorithm employed in our comparison is the one presented in [5] with an excitation density of 20 pulses/subframe. The pulse amplitudes and pulse locations were computed by means of the optimal amplitude method described in [15]. As with the MSLPC case discussed before, the coding of pulse locations was performed by enumerative source coding techniques [14]. Additionally, we quantised the pulse amplitude information by encoding the maximum absolute value per frame and the individual pulse amplitudes as a fraction of the maximum value. As the table shows, the MSLPC coder yielded substantial improvement over the MPLPC coder for all test samples. Our informal listening tests also confirmed the superior quality of the MSLPC output.

Table 2: SNRSEG performances of various audio coders at 64kbit/s

Audio	Coders		
	MPLPC	MSLPC	MPEG–II
Electric instrument	23.52	25.91	26.49
Oboe + piano	22.99	26.79	28.69
Orchestra	22.80	25.51	21.88
Voice	22.38	25.12	24.58

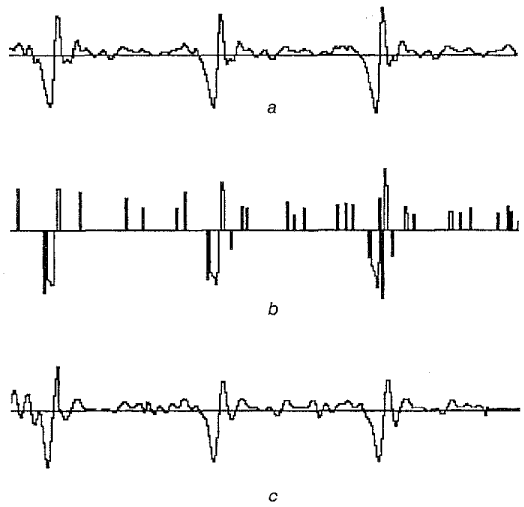

Fig. 4 *Residual waveforms*
a Original
b MPLPC
c MSLPC

Among the reasons for success, we found that a sinusoidal excitation model can more closely match the intrinsic natures of actual residual signals. This is illustrated for a typical audio segment in Fig. 4, where the residual waveform and the corresponding multipulse and multisinusoid modelled versions are included for purposes of comparison. More importantly, from a psychoacoustic perspective, the MSLPC coder has the advantages of using MNR-adapted bit allocation to increase quantiser resolution for perceptually important sine waves. To elaborate further, we also included the performance of the well-established ISO/MPEG 64 kbit/s audio coding system [16]. For layer II at a sampling rate of 32kHz, the MPEG system employs a fil-

terbank to create 24 critically sampled representations of the input signal, which are then quantised using adaptive block companding under the control of the estimated masking threshold. As listed in Table II, MSLPC and MPEG produced comparable performances, with perhaps a slight advantage going to MPEC. Informal listening tests also indicated that the MSLPC output was indistinguishable from that of the coder as standardised by ISO/MPEG-Audio.

5 Conclusions

In this paper, we first emphasised the importance of matching LPC excitation representation to the pulselike natures of residual spectra. This was done by using a sum of sine waves to approximate the excitation waveform, rather than using spectrally flat excitation signals as in multipulse LPC coders. One enhancement that further improves the output quality is the use of mask-to-noise ratio in designing adaptive quantisation for sinusoidal amplitudes. It was found that sinusoidal representation provides an ideal framework for incorporating estimated masking thresholds into noise spectral shaping. Simulation results suggested that the use of a sinusoidal excitation model combined with perception-oriented bit allocation allows implementation of an LPC-based audio coder that delivers high quality at the rate of 64 kbit/s.

6 Acknowledgment

This work was supported by the National Science Council, Taiwan, ROC, under Grant NSC85-2221-E009-029.

7 References

1 EBERLEIN, E., GERHAUSER, E.H., and KRAGELOH, S.: 'Audio codec for 64 kbit/sec, ISDN channel – Requirements and results.' Proceedings of ICASSP, 1990, pp. 1105–1108
2 JOHNSON, J.D.: 'Transform coding of audio signals using perceptual noise criteria', *IEEE J. Sel. Areas Commun.*, 1988, pp. 314–323
3 THEILE, G., LINK, M., and STOLL, G.: 'Low bit-rate coding of high quality audio signals.' Proceedings of the AES Convention, 1987 (Preprint 2432)
4 LIN, X., SALAMI, R.A., and STEELE, R.: 'High quality audio coding using analysis-by-synthesis technique.' Proceedings of ICASSP, 1991, pp. 3617–3620
5 SINGHAL, S.: 'High quality audio coding using multipulse LPC.' Proceedings of ICASSP, 1990, pp. 1101–1104
6 VELDHUIS, R., and KOHLRAUSCH, A.: 'Waveform coding and auditory masking' *in* KLEIJN, W.B., and PALIWAL, K.K. (Ed.): 'Speech coding and synthesis' (Elsvier Science, New York, 1995)
7 CHEN, J.H., and GERSHO, A.: 'Real-time vector APC speech coding at 4800 b/s with adaptive postfiltering.' Proceedings of ICASSP, 1987, pp. 2185–2188
8 ATAL, B.S., and SCHROEDER, M.R.: 'Predictive coding of speech signals and subjective error criteria', *IEEE Trans. Acoust. Speech Signal Process.*, 1979, **ASSP–27**, pp. 247–254
9 KROON, P., and DEPRETTERE, E.F.: 'A class of analysis-by-synthesis predictive coders for high quality speech coding at rates between 4.8 and 16 kbit/s', *IEEE J. Sel. Areas Commun.*, 1988, **6**, (2), pp. 353–363
10 CHANG, W.W., and WANG, C.T.: 'A masking-threshold-adapted weighting filter for excitation search', *IEEE Trans. Speech Audio Process.*, 1996, **4**, pp. 124–132
11 MCAULAY, R.J., and QUATIERI, T.F.: 'Speech analysis/synthesis based on a sinusoidal model', *IEEE Trans. Acoust. Speech Signal Process.*, 1986, **ASSP–34**, pp. 744–754
12 GOLUB, G.H., and VAN LOAN, C.F.: 'Matrix computations' (John Hopkins University Press, Baltimore, 1989)
13 PALIWAL, K.K., and ATAL, B.S.: 'Efficient vector quantisation of LPC parameters at 24 bits/frame', *IEEE Trans. Speech Audio Process.*, 1993, **1**, pp. 3–14
14 COVER, T.M.: 'Enumerative source coding', *IEEE Trans. Inf. Theory*, 1973, **IT–19**, pp. 73–77
15 SINGHAL, S., and ATAL, B.S.: 'Amplitude optimisation and pitch prediction in multipulse coders', *IEEE Trans. Acoust. Speech Signal Process.*, 1989, **ASSP–37**, (3), pp. 317–327
16 ISO/IEC Int. Std. 11172–3: 'Information technology – Coding of moving pictures and associated audio for digital storage media up to about 1.5 mbits/s, Part 3: Audio'